Theses and Dissertations | 1. Thesis and Dissertation Collection, all items

1993-06

# Cognitive limitations in coordination in hierarchical information processing structures

## Armbruster, Robert R.

Monterey, California: Naval Postgraduate School

http://hdl.handle.net/10945/24126

# REPORT DOCUMENTATION PAGE

| eport Security Classification: Unclassified | 1b Restrictive Markings |
|---|---|
| ecurity Classification Authority | 3 Distribution/Availability of Report |
| eclassification/Downgrading Schedule | Approved for public release; distribution is unlimited. |

| rforming Organization Report Number(s) | 5 Monitoring Organization Report Number(s) |
|---|---|

| ame of Performing Organization<br>al Postgraduate School | 6b Office Symbol<br>(if applicable) 39 | 7a Name of Monitoring Organization<br>Naval Postgraduate School |
|---|---|---|

| ddress (city, state, and ZIP code)<br>nterey CA 93943-5000 | 7b Address (city, state, and ZIP code)<br>Monterey CA 93943-5000 |
|---|---|

| ame of Funding/Sponsoring Organization | 6b Office Symbol<br>(if applicable) | 9 Procurement Instrument Identification Number |
|---|---|---|

| ress (city, state, and ZIP code) | 10 Source of Funding Numbers | | | |
|---|---|---|---|---|
| | Program Element No | Project No | Task No | Work Unit Accession No |

itle (include security classification) COGNITIVE LIMITATIONS IN COORDINATION IN HIERARCHICAL INFORMATION OCESSING STRUCTURES

ersonal Author(s) Armbruster, Robert R.

| Type of Report<br>ster's Thesis | 13b Time Covered<br>From        To | 14 Date of Report (year, month, day)<br>June, 1993 | 15 Page Count<br>126 |
|---|---|---|---|

upplementary Notation  The views expressed in this thesis are those of the author and do not reflect the official policy or position he Department of Defense or the U.S. Government.

| osati Codes | | | 18 Subject Terms (continue on reverse if necessary and identify by block number) |
|---|---|---|---|
| | Group | Subgroup | Command and Control, Hierarchical Information Processing Structures, Tactical Decision Making Under Stress, TADMUS, CHIPS, Cognitive Limitations |
| | | | |

bstract (continue on reverse if necessary and identify by block number)

   In Command and Control, the majority of decisions require the fusion of inputs from a number of subordinate decision-ers, to arrive at a team decision.  Part of the Navy's attempt to address the issue of hierarchical decision making is the tical Decision Making Under Stress (TADMUS) program.  Under this program, the Coordination in Hierarchical Processing ctures (CHIPS) experiment was conducted at the Naval Postgraduate School during May and June, 1993.  The CHIPS eriment is described, and data collected during the experiment used to assess the impact of human cognitive limitations on team ormance.

   Team performance was found to be degraded by increased stress, increased risk and increased feedback to subordinates in hierarchy.  These effects were found to be due to a reduced ability to distinguish between types of contact, rather than use of a optimal decision criterion.

   It was further found that increasing the amount of information available to subordinates increased their ability to inguish between types of contacts, but not by as much as is theoretically possible.  There were also indications that there may n upper limit on the amount of information that can be successfully integrated by the subordinates, beyond which performance ines rather than improving.

| istribution/Availability of Abstract<br>nclassified/unlimited ___ same as report ___ DTIC users | 21 Abstract Security Classification<br>Unclassified | |
|---|---|---|
| Name of Responsible Individual<br>hael G. Sovereign | 22b Telephone (include Area Code)<br>(408) 656-2428 | 22c Office Symbol<br>OR/SM |

| ORM 1473, 84 MAR | 83 APR edition may be used until exhausted | security classification of this page |
|---|---|---|
| | All other editions are obsolete | Unclassified |

COGNITIVE LIMITATIONS IN COORDINATION IN HIERARCHICAL INFORMATION
PROCESSING STRUCTRURES

by

Robert R. Armbruster
Lieutenant, United States Navy
M.A., Cambridge University

Submitted in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS TECHNOLOGY

from the

NAVAL POSTGRADUATE SCHOOL
June 1993

**ABSTRACT**

In Command and Control, the majority of decisions require the fusion of inputs from a number of subordinate decision-makers, to arrive at a team decision. Part of the Navy's attempt to address the issue of hierarchical decision making is the Tactical Decision Making Under Stress (TADMUS) program. Under this program, the Coordination in Hierarchical Processing Structures (CHIPS) experiment was conducted at the Naval Postgraduate School during May and June, 1993. The CHIPS experiment is described, and data collected during the experiment used to assess the impact of human cognitive limitations on team performance.

Team performance was found to be degraded by increased stress, increased risk and increased feedback to subordinates in the hierarchy. These effects were found to be due to a reduced ability to distinguish between types of contact, rather than use of a less optimal decision criterion.

It was further found that increasing the amount of information available to subordinates increased their ability to distinguish between types of contacts, but not by as much as is theoretically possible. There were also indications that there may be an upper limit on the amount of information that can be successfully integrated by the subordinates, beyond which performance declines rather than improving.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# I. INTRODUCTION

## A. BACKGROUND

All complex decisions tend to be made within a hierarchical framework, in which a central decision-maker chooses from options based on the reports of subordinate decision-makers. This structure is to be found in business, government, and the military, where it is a fundamental component of command and control systems. In such systems, the decision-makers may be humans or automata, or hybrid systems of humans assisted by computer-based decision aids.

Of particular importance in the military field is the Distributed Dynamic Decision making (DDD) paradigm proposed in Kleinman, Serfaty, & Luh (1984), in which the decision role is not only distributed amongst geographically separated subordinates, but the underlying attributes on which the decision is to be made are dynamic. In such a system, there are three levels of processing to be considered: the individual decisions of each team member at each time of evaluating the dynamic environment, the aggregation by the individual team member into his[1] final decision, and the coordination within the team that leads to a decision by the central decision-maker. At the first level,

---

[1]Throughout this report, the masculine form of the personal pronoun has been used without exception. No specific gender requirement or bias is to be inferred from this; the more cumbersome combination forms such as he/she have been avoided for the sake of readability, while the systematic alternation of gender has been avoided as having even worse implication of bias than the use of a single form throughout.

the assessment of a dichotomous situation confused by the presence of noise is isomorphic to the paradigm of "Signal Detection Theory" (SDT) (Pete, Pattipati, & Kleinman, 1993:2). As further assessments are made, humans are unable to aggregate information optimally (i.e., in the manner of Patterson and Beach's (1967) "statistical man") because they are limited in their data processing capabilities (Pete, Pattipati, & Kleinman, 1993:1; Mallubhatla *et al.*, 1991). Lastly, the decisions of the subordinates are fused, and a final team decision made. Optimization of this production of a team decision requires more than just the optimization of decision for each individual decision-maker, since optimal team performance requires finding a set of *coupled* operating points (Pete, Pattipati, & Kleinman, 1993:1 & 1993:2; Tank, Pattipati & Kleinman, 1991).

## B. SIGNAL DETECTION THEORY AND STATISTICAL DECISION THEORY (SDT)

Signal-Detection Theory has its roots in psychophysics, with the problem of the detectability of a signal in noise. Since "a major part of detection theory is the application of the theory of decision making to situations in which certain waveforms called *signals* may or may not be added to a random background disturbance called *noise*," (Green & Swets, 1988, p. 7) the terminology and structure of detection theory are derived from Statistical Decision Theory. The two theories have become almost inextricably intertwined. Conveniently, they have the same acronym, and this will be used without distinguishing between the theories.

## 1. Structure and Terminology of the SDT Problem

We introduce here the notation[2] that will be used to identify events and decisions by tracking the events in the detection and report of a signal. Firstly, an observation is made: this observation may be of noise alone, labeled "n," or of a signal added to the noise, labeled "sn." The observer then makes a decision about what was observed: either "yes," a signal was present, denoted Y, or "no," a signal was not present, denoted N. The probability that a signal will be presented is $P(sn)$, and that it will not (i.e., that noise alone will be presented) is $P(n)$. These are referred to as the *prior probabilities* of the events sn and n, respectively, and sum to unity.

There are four possible outcomes:

1. Correct acceptance: the occurrence of sn and Y--also termed a *hit*;

2. Incorrect rejection: the occurrence of sn and N--also termed a *miss*;

3. Correct rejection: the occurrence of n and N--also termed a *non-event*;

4. Incorrect acceptance: the occurrence of n and Y--also termed a *false-alarm*.

A pay-off may be associated with each of the four outcomes. This is a score value, or reward, for the subject, which may be positive or negative. The payoffs for the four outcomes are denoted $V_{sn,Y}$, $V_{sn,N}$, $V_{n,N}$, and $V_{n,Y}$ respectively.

---

[2]The easier, dichotomous terminology of Egan (1975, pp. 6-20) from Signal Detection Theory has been used in preference to the more general notation of Green and Swets (1988, pp.13-20) from Statistical Decision Theory.

The conditional probability of a Y response given an sn event, $P(Y|sn)$, is called the *hit rate*. It may also be written as $P(Hit|sn)$, to emphasize that a hit has occurred. Since the event sn must result in either a hit or a miss, the miss rate is $1-P(Y|sn)$. The conditional probability of a Y response given an n event, $P(Y|n)$, is called the *false-alarm rate*. It may also be written $P(False-Alarm|n)$. Given an n event, the response must result in either a correct rejection or a false-alarm, so the correct rejection rate is $1-P(Y|n)$. Thus it is seen that with the hit rate, false-alarm rate, and prior probabilities, the rates (or probabilities) of all four outcomes are completely specified:

$$
\begin{aligned}
P(sn \wedge Y) &= P(Y|sn)P(sn) \\
P(sn \wedge N) &= [1-P(Y|sn)]P(sn) \\
P(n \wedge N) &= [1-P(Y|n)]P(n) \\
P(n \wedge Y) &= P(Y|n)P(n)
\end{aligned}
\qquad (I\text{-}1)
$$

The observation of the event resulted in a measurement $x$, the value or magnitude of which depends probabilistically on which event occurred. In particular, when the event n occurs, the perceived $x$ is a sample from a probability distribution of noise, which is also called "n." Similarly the event sn gives rise to a perceived $x$ from a different probability distribution, called "sn."

The *posterior probability* of the event sn is written $P(sn|x)$, and is the probability that the event sn occurred, given the evidence $x$. The posterior probability of n is defined similarly, and the two probabilities sum to one. A simple application of Bayes rule gives:

4

$$P(sn|x) = \frac{P(x|sn)\,P(sn)}{P(x|sn)\,P(sn) + P(x|n)\,P(n)} \qquad \text{(I-2)}$$

We now define the *likelihood ratio* as the ratio of probabilities of the observed $x$ resulting from the sn and n distributions:

$$L(x) = \frac{P(x|sn)}{P(x|n)} \qquad \text{(I-3)}$$

Dividing the numerator and denominator of Equation (I-2) by $P(x|sn)P(sn)$ gives

$$P(sn|x) = \frac{1}{1 + \left(\dfrac{1}{L(x)}\dfrac{P(n)}{P(sn)}\right)} \qquad \text{(I-4)}$$

Equation (I-4) demonstrates the relation between the posterior probability of sn and the likelihood ratio: as the former increases from 0 to 1, the latter increases from 0 to infinity. The best estimate[3] of the event that has occurred, given the observation $x$, is sn if, and only if, $P(sn|x) > P(n|x)$, which, from the fact that these two probabilities are complementary, is the same as saying that $P(sn|x) > 0.5$. Alternatively, the odds in favour of sn can be calculated (using Bayes rule) as

---

[3]Here we use the criterion of maximizing correct assessments. For maximization of expected value, see the discussion in the following paragraphs.

$$\frac{P(sn|x)}{P(n|x)} = \frac{P(x|sn)}{P(x|n)} \frac{P(sn)}{P(n)}$$
$$= L(x) \frac{P(sn)}{P(n)} \qquad (I\text{-}5)$$

Consider the case where the prior probabilities of the two events are not equal, or the payoffs of the four outcomes are not symmetrical. A common decision goal in these circumstances is to maximize the expected value of the decision. The decision Y should be made if the expected value of choosing Y, given the evidence $x$ and prior probabilities, is greater than the expected value of choosing N with the same evidence. This is true if

$$\frac{P(sn|x)}{P(n|x)} > \frac{V_{n,N} - V_{n,Y}}{V_{sn,Y} - V_{sn,N}} \qquad (I\text{-}6)$$

Combining equations (I-5) and (I-6) gives the *Likelihood Ratio Test (LRT)*: choose Y if

$$L(x) > \frac{P(n)}{P(sn)} \frac{V_{n,N} - V_{n,Y}}{V_{sn,Y} - V_{sn,N}} \qquad (I\text{-}7)$$

## 2. The Relative Operating Characteristic (ROC)

The distributions sn and n are assumed to overlap to a greater or lesser degree--if they did not, no errors should ever be made. Figure 1 shows two such distributions: these in particular are equivariant normal (Gaussian) distributions, but there is no requirement that they be so. We define $x_c$ as the value of $x$ chosen as a criterion, corresponding to the critical value of $L(x)$. As higher values are chosen for $x_c$, the false-alarm rate will fall; however, the hit rate will also fall, as shown in Figure 2.
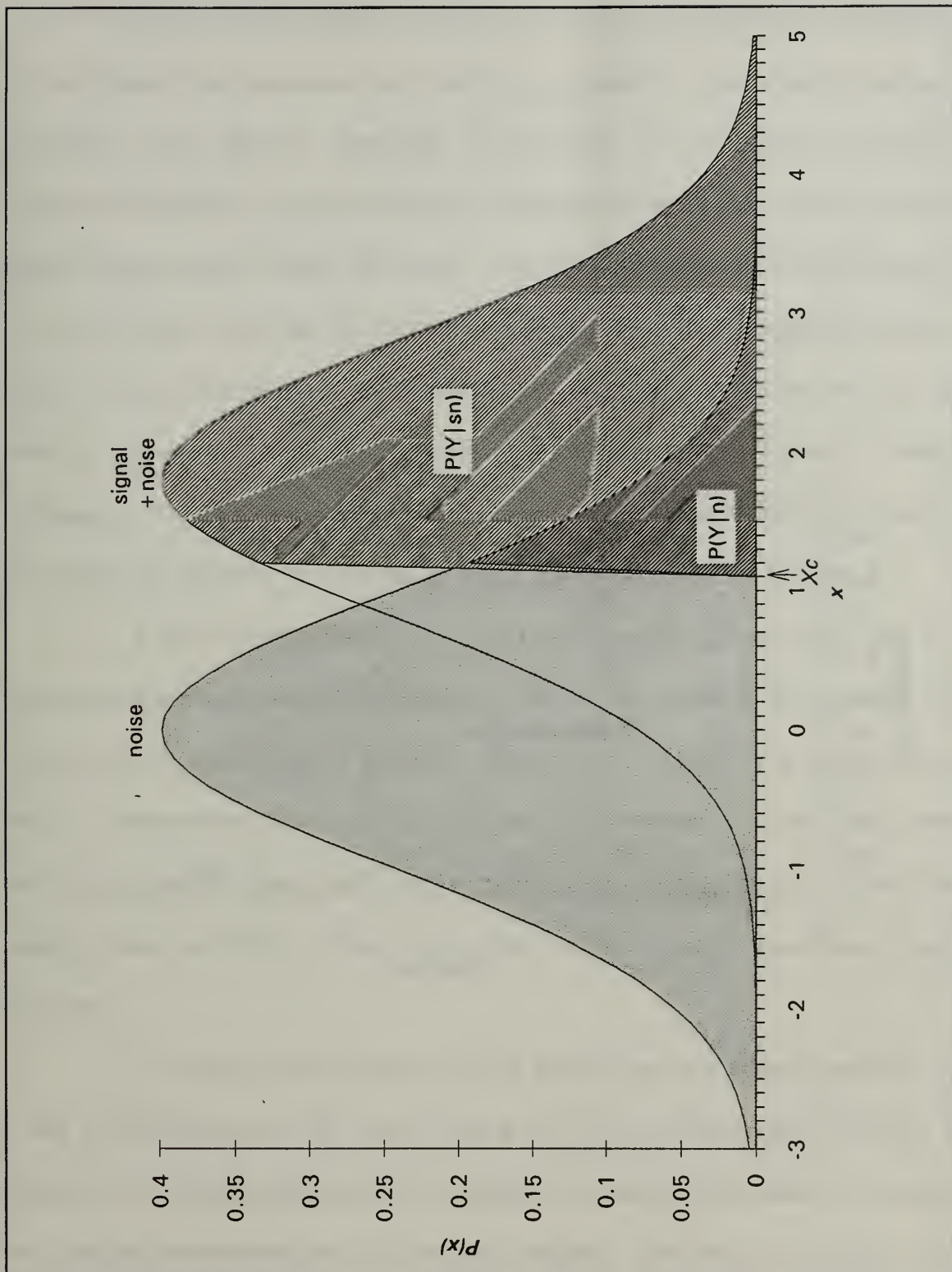
Figure 1: Distributions of Noise and Signal+Noise, with Hit and False-Alarm
Probability Areas Shown
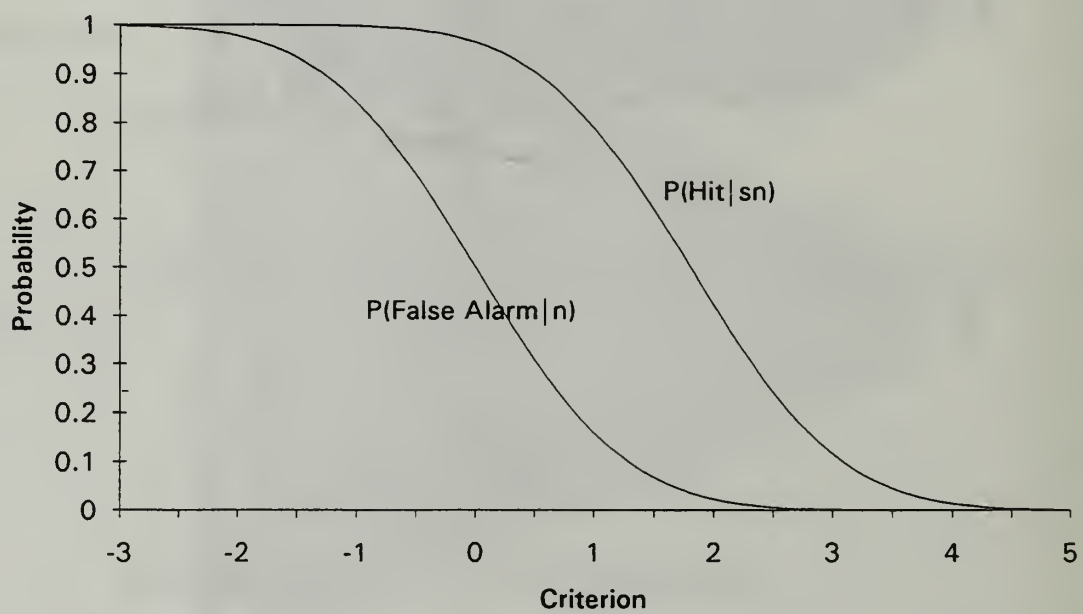
Figure 2: False-Alarm and Hit Probabilities as Functions of Criterion Value

For any given $x_c$ there is a unique hit rate and false-alarm rate. Plotting the hit rate against the false-alarm rate, with $x_c$ as parameter, gives a curve variously described as the *Receiver Operating Characteristic* or the *Relative Operating Characteristic* (ROC). Again, the acronym is the same for either name, and can be used without distinguishing between the names. This curve describes the distinguishability between the sn and n events for the observer, without assuming a particular criterion value. The criterion selected will determine the *operating point* on the ROC of the observer. An example of the ROC generated by the distributions in Figure 1 is shown in Figure 3. The minor diagonal, identified in the figure, is also known as the *chance line* because an observer who guesses randomly would operate along this line.

A ROC that is based on use of the LRT is termed a *proper ROC*. For any given false-alarm rate, the corresponding hit rate on the proper ROC represents the maximum that can be achieved given the probability distributions of n and sn (Egan, 1975). Characteristics of a proper ROC that aid in ROC analysis are that it lies above and to the left of the chance line, and is non-decreasing throughout; i.e., shifting the operating point on the ROC to one with a higher hit rate can never give a lower false-alarm rate.

For families of ROCs that are all the same shape, such as those generated by n and sn distributions that are both normal and have the same standard deviation, a particular curve within the family can be identified by specifying the distance along the main diagonal between the ROC and the minor diagonal. This distance is called $d'$, and is the distance between the means of the n and sn distributions, measured in standard
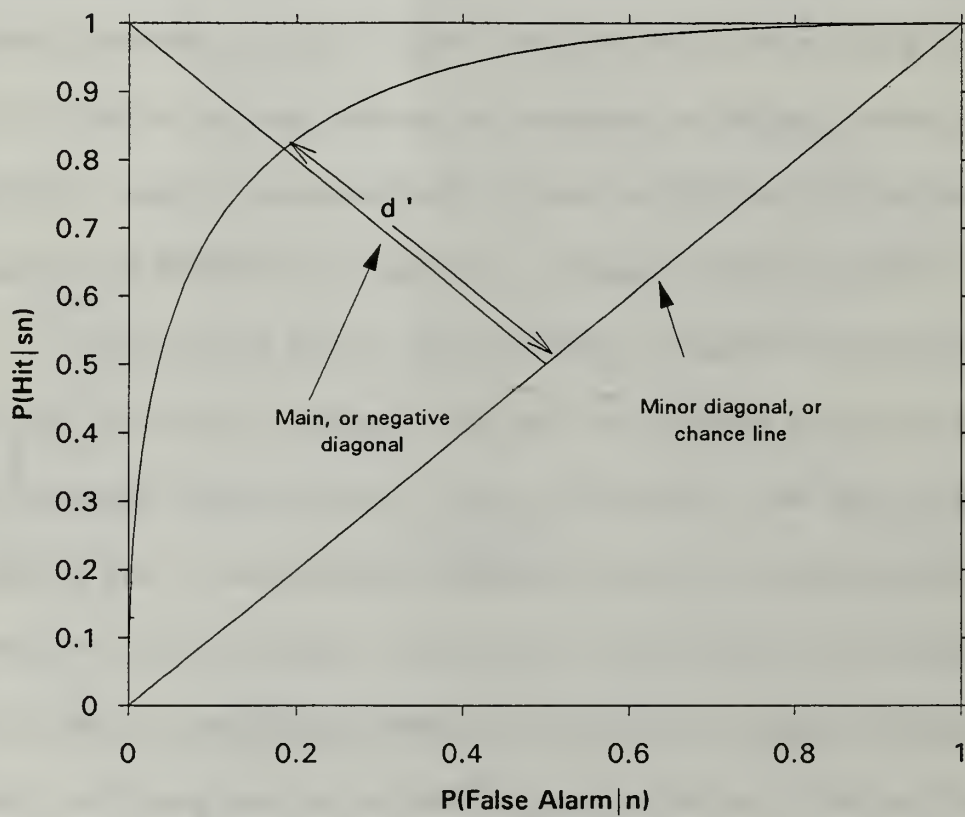
Figure 3: Relative Operating Characteristic (ROC) for the Distributions in Figure 1

deviations of the n distribution:

$$d' = \frac{\mu_{sn} - \mu_n}{\sigma_n} \qquad \text{(I-8)}$$

When the hit rate and false-alarm rate are converted to normal-deviate values,[4] the ROC is a straight line for normal n and sn distributions (Swets, 1986:1 & 2), and is parallel to the chance line when the distributions are equivariant. Thus for equivariant normal distributions, an observed $d'$ can be readily calculated as

$$d' = z_{Hit} - z_{False\,Alarm} \qquad \text{(I-9)}$$

When the distributions are normal but not equivariant, the ROC is not parallel to the chance line in plot of Z-scores, so equation (I-9) must be replaced by

$$d' = \frac{\sigma_n}{\sigma_{sn}} z_{Hit} - z_{False\,Alarm}$$

(Macmillan & Kaplan, 1985).

### 3. Multiple Observations.

When the decision is based not just on one observation, but on several observations, each drawn from the same distribution, the parameters of the SDT problem change. Several different models may be used to describe the change in decision strategy, but in general the results are the same. The more psychophysiological model proposes an integration of the observations, and a decision based on the integral (Green

---

[4]For example, $P(\text{Hit}|\text{sn})$ would be replaced by $z_{Hit}$, where $P(\text{Hit}|\text{sn}) = 1 - \Phi(z_{Hit})$. $\Phi(z)$ is the cumulative standard normal distribution.

& Swets, 1988, p. 238, and Swets, et al., 1959). The average of $N$ samples from the same normal distribution will have the same mean as the underlying distribution, but the variance will have been reduced by the factor $N$. Thus, from equation (I-8), $d'$ will have been increased by the factor $\sqrt{N}$. The more statistical model states that the likelihood ratio for the $N$ observations together is the product of the likelihood ratios of the individual observations, assuming that they are independent (Egan, 1975, p. 77). This new likelihood ratio can then be used in the LRT. The significance of this dichotomy with respect to the observed data is discussed in Chapter V.

## C.  COGNITIVE LIMITATIONS

Thus far we have discussed how an optimal observer, with knowledge of prior probabilities of the two possible events, knowledge of the probability distributions that could have given rise to the observed datum, knowledge of Bayes law and its ramifications, and the ability to process all this knowledge--including calculation of values with the Gaussian probability density function, can formulate his rule for optimum performance. Evidently, humans *can not* respond in this way. Even in gambling, when the distributions and their implications have been well studied by practitioners with a strong interest in doing well, optimum performance cannot be achieved immediately. There exist a series of card counting schemes for Blackjack, increasing in expected pay-off and cognitive demand, even the simplest of which takes considerable mental effort

and practice, amounting to the dedication of one or two months to the study of the method.[5]

The human subject has cognitive limitations, which should be accounted for in any theory on decision-making. Pitz summarizes the types of limitations:

> Failures to respond consistently might be traced to one of two sources. First, there may exist limitations on the kind of information processing of which the person is capable. ... A second source of errors may be the problem solving strategies that a person brings to a task. Such errors are not due to fundamental limitations on information processing capacity, but rather to the strategies that people use in approaching the task. (Pitz, 1980, p. 78)

Perhaps the most fundamental limitation on information processing is the general inability of the mind to handle more than about seven chunks in short-term memory, or seven channels in discrimination tasks. These limitations are distinct: the first is the most items that can be recalled after a brief interval when rehearsal is prevented; the other is the maximum number of different stimuli that can be distinguished. The fact that in both cases the limit is the "magical number seven" is probably coincidental. (Miller, 1956)

Human subjects are, in general, poor at judging statistical parameters of distributions, as described by Peterson & Beach (1967). Proportions are assessed well, particularly when they are not extreme. There exists conflicting evidence about whether high proportions tend to be over-estimated or under-estimated. Estimation of means is also accurate, although accuracy diminishes with increasing sample size and variance.

---

[5]Interview between W. Snow, Maj., USAF, Naval Postgraduate School, Monterey, CA, and the author, 3 May, 1993.

Judgments of variances are poor: reported values are not related to squared deviations, instead being based on much lower powers of the deviations, and are strongly influenced by the mean of the sample--the higher the mean, the smaller the reported variance. When making inferences about populations, such as whether a sample comes from one distribution or another, subjects are uniformly conservative in assessing the value of the evidence of the information provided by the sample (see also Tversky & Kahneman, 1974). Additionally, aggregation of evidence from several samples is extremely poor, being very conservative.

Many biases evident in decisions are described by Tversky and Kahneman (1974), of which only one will be listed here. Despite knowledge of the prior probabilities of different outcomes, subjects did not always use this knowledge to modify their estimate of posterior probabilities. For example, subjects were asked to assess the probability that an individual (from a group of engineers and doctors) was an engineer, given a description of him. The subjects were told proportions for engineers and doctors within the base group, thus providing them with prior probabilities. Yet this prior probability had no influence on the subjects' assessments. Even when the description had no distinguishing information at all, the response of subjects was the same whether the prior probability of engineers was 0.3 or 0.7.

Finally, the biases of recency and primacy, which were originally described in relation to memory (Glass, Holyoak, & Santa, 1979, pp. 148-149), are also to be found in decision experiments. Primacy is the excessive influence on the decision of data presented at the start of a trial; the subject anchors on the initial data, and then uses

insufficient adjustment to account for later data (Tversky & Kahneman, 1974). Recency is the strong influence on the subject's decision of the most recently occurring data.

## D.    HIERARCHICAL PROCESSING

Actual decisions in complex environments are generally made by a team, which is usually arranged hierarchically: in the simplest case as a central decision-maker, provided estimates by subordinates.  Each member of the team has access to information that is probabilistically related to the group decision.    The local ROC of each subordinate, which expresses his hit and false-alarm rates with respect to <u>his</u> own task, can be extended to a perceived ROC, which expresses his expertise with respect to the team task.    The behavior of the team can then be summarized by a team ROC:    the relationship between hit and false-alarm rates of the team as a whole.    The optimal behavior of the team requires each team member to adapt his strategy to the expertise of the other members, the team goal, and the relationship between his local information and the team goal.  (Pete, Pattipati, & Kleinman, 1993:1)

## E.    OBJECTIVES

The experiment analyzed in this thesis was conducted at the Naval Postgraduate School during May and June, 1992, and titled *Coordination in Hierarchical Information Processing Structure* (CHIPS).  The goal of the experiment was to validate normative model predictions about hierarchical decision-making in a dynamic, distributed scenario. Since the hierarchical aspects of the experiment are being fully analyzed elsewhere

(principally at ALPHATECH, Inc. and the University of Connecticut), they will not be considered here. The goal of this thesis is to examine the subordinate decision-making process, and attempt to describe cognitive limitations that lead to performance below the optimal achievable.

## II. DESIGN OF THE EXPERIMENT

### A.    THE CHIPS PARADIGM

The CHIPS experiment was designed as a functional simulation of the Anti-Air Warfare components of a surface ship's Combat Information Center.  A team of subjects is organized as four decision-makers, one of whom is the Tactical Action Officer (TAO), who leads the team and makes the team decision.  Supporting him are three subordinates, designated as the Identification Supervisor (IDS), the Tactical Information Coordinator (TIC) and the Electronic Warfare Supervisor (EWS).  The goal of the team is to determine whether a target of interest is hostile or neutral, before the time limit of the trial is exceeded.

The subjects are all presented with a display simulating a radar picture, on which aircraft icons appear and move.  One icon is designated the target of interest by the computer running the simulation.  It is clearly distinguishable by its unique icon.  The subordinates can probe the target, in order to measure certain attributes about the target. Each subordinate measures a different attribute of the target.  After a ten-second delay (to limit the probe rate) the result of the probe may be read by the subordinate, or the target may be probed again.  When the subordinate chooses to read the results of his probes, he opens a window on his display, in which are presented the results of his probes, up to a maximum of the five most recent.  If a probe completes its ten second wait while the window is open, its information is _not_ displayed until the next time the

17

subordinate opens the assessment window. The information presented to the subordinate is corrupted by "noise," which is simulated by adding to the parameter a random value[6] before making it available for display. On the basis of the information displayed, he must make an assessment about the nature of his parameter, which is a dichotomous choice. No further probes may be initiated until the assessment has been made. Once the subordinate makes an assessment, this assessment becomes available to the TAO. The subordinate assigns a confidence to his assessment, on an integer scale of 1 to 3, 1 being the lowest confidence.

Each subordinate is attempting to determine whether his own parameter is indicative of a hostile target or a neutral target. The TAO may open a window on his display on which the most recent assessments and confidences of the subordinates is displayed. As with the subordinates, new information received while this window are open is not displayed. On the basis of the information reported by the subordinates, the TAO makes an overall determination of whether the target is hostile or neutral. At any point during the trial he may designate his assessment as final, which ends the trial. The actual state of each attribute and the target as a whole are displayed to the team at the completion of the trial. The TAO was prompted (by the computer) to make an assessment every 30 seconds.

The state (hostile or neutral) of the three attributes of the target are independent. If two or three of the three are hostile, then the target is hostile. Otherwise (i.e., two

---

[6]The EWS information is corrupted differently. Full details for each subordinate are presented in the discussion of individual subordinate roles in section D.1.

or three attributes are neutral) the target is neutral. Thus, the actual state of the target is determined by a majority rule of its actual attributes.

## B. EXPERIMENTAL SETUP

### 1. Physical Setup

The simulation was conducted on a network of four Sun workstations, running a version of the DDD-II simulator specifically developed for the CHIPS experiment. Each subject's area contained a graphics display screen, keyboard, mouse, and intercom headset. The areas were separated within a single room, to attempt to provide some isolation, and prevent un-monitored communication between the subjects. The keyboard was not used except to start each trial. All probes and assessments were accomplished with the mouse. Subjects were permitted to use pencil and paper, although this was not provided, and none took advantage of the permission. Subjects were not permitted to use calculators.

The intercom system connected all the subjects together--when any one spoke, he was heard by all. Verbal communication was allowed between all the subjects without constraint, with the exception that for some trials the TAO was not allowed to brief the subordinates on his assessment of the target (see section D.2). To assist in the conduct of the experiment, and record frequency of types of verbal communication, each subject was monitored throughout the experiment by an observer.

## 2.    Test Subjects

The subjects were 24 students from the Joint Command, Control and Communications curriculum at the Naval Postgraduate School, in Monterey, California. The 23 military officers and one National Security Agency employee were divided into six teams of four subjects, based on scheduling constraints. Assignment of function within the team was made by the team members themselves.

Training on the conduct of the experiment consisted of three stages. First, the subjects were provided with written material outlining the background behind the scenario assumed by the experiment. Second, a one hour training session[7] reiterated the written material, stressed the roles of each subordinate and the relation between subordinate task and team goal, and answered any questions that the subjects had. Finally, 24 trials were conducted in two, one-hour training sessions on the actual hardware, during which time the subjects were coached on the mechanics of their tasks by the observers. This exposure allowed the teams to discuss and decide on tactics to be employed, particularly regarding communications.

---

[7]Details of experimental design were intentionally omitted. The variability of the distributions sampled to impose noise on the subordinates measurements was not briefed to the subjects (even the observers had to resort to examining the source code to get a definitive answer, since the various sources of information available to them differed in their details). The statistical design of the experiment, including the existence of distractor trials, was not briefed. However, the prior probabilities of hostile and neutral parameters was, in general terms, made clear to the subjects as being 0.5 each.

## C.   ASSUMPTIONS

Two assumptions were made about the structure of the experiment. Firstly, it was assumed that the training sessions were sufficient to preclude observing a learning-curve effect in the experimental data. Initial trials by the observers indicated that the game can be well-learned in an hour. By distributing the two hours of training over two one-hour sessions the training was very effective, with minimal loss of skill between completion of training and the experimental session. Subject perception of completeness of training was measured with a questionnaire. It was found that, given the average gap of five days between completion of training and the experimental session, the first one to three experimental trials involved some relearning of the necessary skills. Future experiments should have about five refresher trials, known as such by the subjects, at the start of the experimental session.

Secondly, it was assumed that the subjects were willing and enthusiastic, and that the data were therefore not affected by half-hearted guessing on the part of the TAO or his staff. A reward was promised for the top-performing team to help to alleviate this concern. During the experiment, two TAOs accidentally recorded a few assessments as final inadvertently, and were surprised to see that they had made a successful guess. One of them continued this  with some further trials, and was in general successful at guessing. However, his team achieved the lowest score of the three.

Several assumptions were made about the data, which are described in detail in Chapter IV when the results are discussed.

## D.    EXPERIMENTAL DESIGN

### 1.    Subordinate Roles

#### a.    *Identification Supervisor (IDS): Determination of Radar Cross-section*

The team member performing this task received two measures as the result of his probe of the contact.  These two measures, nominally representing a height and width, were first multiplied together, then the product evaluated.  Large contacts had a radar cross-section with a mean of 60, which was an indication that the target was neutral;[8] small contacts had a radar cross-section with a mean of 40, which was an indication that the target was hostile.   In making their determinations team members were cautioned to remember that they are receiving noisy measurements.  It was also to their advantage to probe for measurements more than once to help them "average out" noise in order to determine the truth.  To provide the measurements, two random processes were employed.  The first process selected a number from either a distribution of small or large targets, as designated for the scenario.  The small distribution had a mean of 40, the large had a mean of 60, and each had a standard deviation of 3.  Like all the Gaussian distributions used in CHIPS, the distribution was truncated at two standard deviations[9].  The second random process determined how much noise was

---

[8]Note that size was only one indication: if size indicated neutral, but the other two indications indicated hostile, the target was hostile.  Despite the attempt of the observers to make this perfectly clear during the training, it was still not understood (and even argued) by some subjects during the experiment.  See section D, p. 37.

[9]Examination of the source code reveals that this is achieved in the following way: If the result of the random number draw (from a generator that produced normally distributed random numbers) gave a result more than two standard

associated with the measure, by adding a random number from a Gaussian distribution with a mean of zero and a standard deviation of 20. To provide the two measures expected by subjects, the noise-corrupted single number was factored into two numbers. This was accomplished by selecting a random number within the bounds of the single number and dividing the single number into the random number to produce two numbers. The product of these two numbers equals that of the single noise-corrupted number. The numbers generated in this fashion were provided to the subjects as measures of radar cross-section height and width.

### b. *Tactical Information Coordinator (TIC): Determination of Altitude Rate of Change*

The member of each team performing this role was provided with a single altitude estimate (in feet) and time of measurement with each probe. When the probe window was first opened an initial altitude measure and the new altitude measure were displayed. In this way a determination of altitude rate of change could be made starting with the first probe. Thereafter, a short history of altitude measures was maintained and displayed along with the new measure. The rule of the altitude rate of change described to the team member stated that if the altitude of the contact appeared to decrease by about ten feet per second the contact was assumed to be flying approximately level, which was an indication that the target was neutral. However, the favoured attack profile of the enemy was known to be a descent of about 20 feet per

---

deviations above or below the mean, the result was replaced with the mean plus or minus two standard deviations, respectively.

second. To produce the altitude estimates, two Gaussian distributions were maintained. One distribution had a mean descent rate of ten feet per second and the other had a mean descent rate of 20 feet per second. Each distribution had a standard deviation of 3 feet per second. For a given condition, the appropriate distribution was sampled and a noisy descent measure selected. Based on the true previous altitude, the time elapsed, and the new noisy rate, a new, true altitude estimate was computed. Yet another Gaussian distribution was sampled (mean of zero, standard deviation equal to 90 feet) to determine the amount of noise to be associated with the altitude estimate.

### c. Electronic Warfare Supervisor (EWS): Determination of Radar Emission

A probe issued by the team member performing this role returned a seven-bit binary number. This measurement was compared to the known radiation signature of hostile aircraft[10] to see if the contact was radiating or the sensor was merely reporting noise. The known signature was another seven-bit number that remained constant throughout the experiment. To provide radar signature estimates, the true signature was corrupted by noise. The noise corruption was achieved by deciding whether each bit of the true signature was to remain unchanged or be flipped by noise. If the radar was on, then each bit of the true signature was conditioned by a Bernoulli trial where the probability of being correct was 0.7. That is, 70% of the time the bit remained unchanged and 30% of the time the bit was flipped to the other state (i.e., 0

_____

[10]Not all hostile aircraft radiated, and some neutral aircraft did radiate: see note 8 on p. 22.

24

becomes 1 or 1 becomes 0). If the radar was not on, then each bit was generated randomly, conditioned by a Bernoulli trial with equal probability of a one or a zero.

## 2. Independent Variables

Three independent variables were manipulated in the experiment: information structure, risk, and stress.

### a. Information Structure, or TAO Update

This variable had two states: update or no update. In the update condition, the TAO was directed to tell the team, over the intercom, what his assessment of the target was at least every thirty seconds (i.e., when he logged an assessment on the computer). In the no update condition, the TAO was not allowed to inform the subordinates what his assessment was at any time during the trial (communication between trials was unrestricted and not logged by observers). Two TAOs, while not briefing the subordinates on their assessments during "no update" trials, did frequently report the individual indications of each subordinate as shown in the display window on their screen (e.g., "I'm showing large, level, radar on").

### b. Risk

This variable also had two states: high risk and low risk; the values in the payoff matrix depended on the risk state, in the following way: in low risk trials, a correct decision resulted in a score of 0; either a false-alarm or a missed detection (the two types of incorrect decision) resulted in a loss of one point; in high risk trials, the payoff matrix differed only in that the cost of a missed detection was five points, while

25

the cost of false-alarms remained one point. In high risk trials the targets were said to be carrying chemical or biological weapons, rather than the conventional weaponry of low risk trials. Thus a perfect score would be 0, and as performance became worse, the score became more negative. The team task was to maximize (make as close to 0 as possible) their total score accumulated across all the trials.

Initial trials conducted to train the observers on the mechanics of the experiment indicated that the risk factor was not even noticed by subjects. In order to give the subjects the incentive necessary to make the risk meaningful, the scores (0, -1 or -5) were accumulated over all trials. The team that performed best was to be rewarded. The number of points lost for each type of error was made known to the subjects, so that there could be no question but that the costs assumed by the model were the same as those perceived by the subjects. The total scores of teams who had already completed the experiment were posted, to provide extra incentive. This may not have been the best policy, since at least one team tried hard until they had already lost more points than the team composed only of women, at which point the TAO lost further incentive for good performance.

### c.  *Time Pressure, or Stress*

Stress was manipulated by varying the time available in which to make the assessment of the target. There were three levels of stress: low stress trials lasted just over three minutes, medium stress trials just over two minutes, and high stress trials just over one minute. The time remaining was prominently displayed on each players screen. If no final assessment was made before the time expired, points were lost as for

a missed detection regardless of target classification. This only occurred in three of the 192 trials.

### 3. Scenarios and Statistical Design

The two values of size (large and small), two values of descent rate (level and descending), and radar emission (on and off) combine to give eight possible combinations, which are shown in Table I. An overall assessment of neutral is correct whenever two or more of the measures indicate neutral.

TABLE I: ALL POSSIBLE COMBINATIONS OF TWO LEVELS EACH OF RADAR CROSS SECTION, ALTITUDE RATE, AND RADAR EMISSION

| Classification | Radar Cross Section | | Altitude Rate | | Radar Emission | |
|---|---|---|---|---|---|---|
| 1  Neutral | 1 | large | 1 | level | 1 | off |
|  | 1 | large | 1 | level | 2 | radiating |
|  | 1 | large | 2 | descending | 1 | off |
|  | 2 | small | 1 | level | 1 | off |
| 2  Hostile | 1 | large | 2 | descending | 2 | radiating |
|  | 2 | small | 1 | level | 2 | radiating |
|  | 2 | small | 2 | descending | 1 | off |
|  | 2 | small | 2 | descending | 2 | radiating |

While all situations were presented to the subjects, those corresponding to the top line and bottom line of the table were not used in the statistical analysis of team performance, in order to examine only cases with a high level of ambiguity. This means that a mistake by any one subordinate could cause the team to make the wrong decision. There were two trials from the top and bottom lines in each set of 32 trials presented to the team.

27

The two levels of information structure, two levels of risk, and three levels of stress combine to determine 12 experimental conditions. Each condition was presented twice, once with a neutral target, and once with a hostile target,[11] giving 24 trials across which the independent variables were balanced. The remaining eight *distractor trials* consisted of the two trials for which all three subordinate indications were the same (either all hostile or all neutral) and six more random scenarios. The 24 balanced experimental trials were the only trials analyzed when examining team performance, in general. When analyzing subordinate performance individually, however, all 32 trials have been used. This was necessary to increase the amount of data available. The effect of the independent variables was not analyzed for subordinates. Therefore there was no necessity for them to be balanced for each team. Balance would have been preferable, since there is evidence that they did have an effect.

4. **Measures**

Three general types of dependent variables were collected during the experiment: data recorded automatically by the DDD simulator, data manually recorded by observers, and self-report measures derived from questionnaires completed by the subjects. Of the first category, every probe and assessment during the progress of the game was recorded in an Event Log file. The most important values from each trial were summarized in a Dependent Variable (DEP) file. The format of each file is described in Chapter III.

---

[11]See section C.2, p. 36 for the one exception.

Two types of observation form were used to record observations during each trial, one for subordinates and one for the TAO. On these forms instances of supplying or requesting information were tallied by type of information and the player addressed. In addition, team bolstering comments and action requests were tallied, as were cases where the TAO provided updates to the subordinates on his hostility assessment when he was not supposed to, or failed to provide updates when he was required to. The data from the forms and questionnaires have not been used for the present study, so examples of these are not shown.

Three questionnaires were used. During training each subject indicated on a scale how well trained he felt after every three trials. After each experimental trial an adaptation of the NASA TLX bipolar rating scale was used to measure perceptive workload and stress. After each group of eight trials, a Post-Session questionnaire was used to measure subject perception of performance. At the end of the experimental session, each subject completed a Debriefing questionnaire.

Two measures of effectiveness have been used in performing the data analysis: proportion of correct assessments and $d'$, the distance from the chance line to the ROC. The proportion of correct assessments is affected not only by the ability to distinguish between the n and sn distributions (i.e., the distributions which produce hostile and neutral parameters), which is measured by $d'$, but also by the effectiveness of the operating point on the ROC (i.e., the critical value against which the likelihood ratio is compared). Since the teams were operating to minimize cost (achieve the best score, which is the score nearest to zero), the cost would perhaps have been a better

measure of the optimality of team performance. However, the distribution of scores is not well defined, so that the statistical tests that could be used would be less powerful[12]. Additionally, for the low risk trials, the goals of maximizing score and maximizing proportion of correct assessments are the same, since the payoff matrix is symmetrical. The utility of the measure $d'$ is discussed further at the start of Chapter IV.

---

[12]For a further discussion of the utility of score as a measure, see note 16 on p. 46.

# III. DATA DESCRIPTION

## A.   EXAMPLE OF RAW DATA

Appendix B presents samples of the raw data that was automatically collected by the CHIPS software.  These data consist of two files which were created for each trial: the Event Log File and the Dependent Variable File.  In addition to the data which were automatically generated, each subject was observed during every trial, and answered questionnaires.  The observations of the subjects gathered information about:

1.   Information transfers;

2.   Information requests;

3.   Bolstering comments;

4.   TAO failures to comply with update requirements.

Questionnaires were completed by the test subjects after each trial, after each session (8 trials), and after the entire experiment (32 trials).  The data provided by the questionnaires consisted of:

1.   Self evaluations;

2.   Teammate evaluations;

3. Team evaluations including:

- Coordination
- Amount of communication
- Communication discipline
- Helpfulness of other team members

4. Stress level evaluations.

## B. DATA CODING SCHEME

### 1. Dependent Variable File

The example provided in Appendix B may be interpreted with the following notes.

#### a. *Experiment Condition*

The experiment condition is a five digit number describing the scenario, interpreted as shown in Table II.

#### b. *Probe Rate*

The rate at which the subject probed the target, in probes per second.

#### c. *Other Codes*

Neutral: 1

Hostile: 2

Low confidence: 1

Medium confidence: 2

High confidence: 3

| Digit | Variable | Coding | | | |
|---|---|---|---|---|---|
| 1st | Information Structure | 1:  No update | | | |
| | | 2:  Update | | | |
| 2nd | Risk | 1:  Conventional (miss weighted same as false-alarm) | | | |
| | | 2:  Chemical (miss weighted five times as heavily as a false-alarm) | | | |
| 3rd | Stress | 1:  Low -- 3 minute trial | | | |
| | | 2:  Medium -- 2 minute trial | | | |
| | | 3:  High -- 1 minute trial | | | |
| 4th | True Classification | 1:  Neutral | | | |
| | | 2:  Hostile | | | |
| 5th | Target Parameters | Value | Size | Descent Rate | Radar |
| | | 3 | Small | Level | Off |
| | | 2 | Large | Descending | Off |
| | | 3 | Large | Level | On |
| | | 3 | Small | Descending | Off |
| | | 6 | Small | Level | On |
| | | 6 | Large | Descending | On |
| | | 7 | Large | Level | Off |
| | | 8 | Small | Descending | On |

## 2.    Event Log File

The log file is meant for internal usage, and therefore the meanings of the numbers in this file are different from those in the experimental description.  Two types of log messages are recorded in the CHIPS experiment:

1.    Probe: code 2010

2.    Fusion/Assess: code 2013

The formats of these messages are as follows:[13]

### a. *Probe*

The C language code that generates a probe entry is as follows:

```
fprintf(logfp, "%d %d   %lf\n", dm, message_code, current_time);
fprintf(logfp, "%d %d\n",platform_id, END_NOTIFIER);
fprintf(logfp, "%f %f %f\n",weapon[0],weapon[1],weapon[2]);
fprintf(logfp, "%d %d   %f %f\n", task_number, dm_flag, delay, expertise);
fprintf(logfp, "%d\n", message_flag);
```

where:

dm - the decision-maker who issues the command.
message_code - equals 2010 for "probe"
current_time - the time when the command is issued
platform_id - not used in CHIPS
END_NOTIFIER - not used in CHIPS
weapon[i] - not used in CHIPS
task_number - not used in CHIPS.
dm_flag - not used in CHIPS.
delay - time delay of "probe", always 10 seconds in CHIPS.
expertise - not used in CHIPS.
message_flag - not used in CHIPS.

### b. *Fusion/Assess*

The C language code that generates a fusion (assessment) entry is as

follows:

```
fprintf(logfp, "%d %d   %lf\n", dm, message_code, current_time);
fprintf(logfp, "%d %d %d %d %d\n", from_dm, to_dm, task_id, classid, confidence);
fprintf(logfp, "%f %f %f\n)",attributes[0],attributes[1],attributes[2]);
fprintf(logfp,"%d\n", flag);
```

---

[13]E-mail conversation between Anlan Song, University of Connecticut, and the author, 15 March, 1993

where:

dm - the decision-maker who issues the command.
message_code - equals to 2013 for "Fusion/Assess"
current_time - the time when the command is issued
from_dm - equals to dm in CHIPS .
to_dm - the decision-maker to whom the message is sent
  = 0, subordinates send message to DM0 (the leader)
  = 4, DM0 (the leader) logs information to the system.
task_id - not used in CHIPS.
classid - estimated target identification
  =0 neutral
  =1 threat
confidence - the confidence level of the decision.
attributes[i] - not used in CHIPS.
flag - not used in CHIPS.

### 3. Questionnaires and Observation Forms

The data from these forms were not used here for analysis, so the coding scheme is not explained here.

## C. DATA PROBLEMS

### 1. Event Log File

There were an insufficient number of runs for each team to be able to analyze $d'$ for each team at each number of probes.

There was an unbalanced distribution of targets for the IDS and EWS with low stress. Specifically, team E IDS had no small targets at low stress, and teams A and B had very few radiating targets at low stress. Consequently, it was not possible to calculate $d'$ accurately for the IDS and EWS by team based only on low stress runs. Combined data were used from all stress levels which limited the number of probes for

35

which useful data existed, since there were only about four probes during high-stress trials, and eight probes during medium-stress trials.

Not many players made more than seven probes, even in low stress runs. Team F, in particular, made very few probes in each trial.

Not infrequently, even when a significant number of data existed, there were either perfect hit rates or false-alarm rates (1.0 or 0.0, respectively), which could not be transformed to Z-scores. The hit or false-alarm rate was replaced with 0.99 or 0.01 respectively in these cases.

Teams A and B EWS had perfect hit rates, but very high ($>0.75$, consistently) false-alarm rates, indicating a quite extreme criterion in use; consequently the calculated $d'$ is prone to inaccuracy.

## 2. Dependent Variable File

It was found that with the software provided by the University of Connecticut, some files would be over-written by subsequent trials that had the same scenario. The scripts directing execution of the CHIPS software were modified to remove this problem.

The data could not be fully balanced with the scenarios as provided. Specifically, team B was given no neutral target in the low stress, low risk, no update condition. Two trials with hostile targets in this condition were used for team B to balance the data.

## D. OTHER PROBLEMS

Of the 24 test subjects (6 teams of 4 players each), one player received no classroom training.

Two of the six teams (A & D) received the first hour of hands-on training prior to classroom training. Team A completed 15 practice trials during this first hour of hands-on training, then the remaining nine practice trials during the second hour.

Due to a power failure, one team (F) received only eight trials during their first hour of hands-on training. Subsequently, team F was scheduled for and received the four remaining trials from the first hands-on training session on a second day, and the remaining 12 practice trials on a third day.

Some of the team training sessions were conducted in a hardware "consolidated" environment (i.e., the stations were not in separate bays) while others were conducted in the experimental format of separate subordinate and TAO locations.

Several of the teams maintained a distinct level of confusion concerning the categorization requirements of an unknown contact based on the three technical or sensed parameters. While it was clearly briefed and demonstrated during classroom and hands-on training that only two of the three ground truth parameters had to fit the hostile criteria for the contact to be considered hostile, some players maintained a belief that it required three of the three to be hostile. Additionally, some players felt that if their fellow subordinates had a high confidence in the identity of the target, then it was incumbent upon them to "fit" their classification to agree with their comrades. Interesting as well, one of the players felt that if two of the three subordinates were

37

correct in their assessment, then regardless of the third subordinate or the TAO's final assessment, the team would be scored as correct. All the TAOs, however, knew correctly the definitions of hostile and neutral contacts.

# IV. ANALYSIS

The analysis of the experimental results falls into two groups. First, there are results of team performance, based largely on the summary data available in the Dependent Variable File. These results generally use the proportion of correct assessments. Secondly, the performance of the individual subordinates (by role) was examined from an SDT view-point, to look for influence of cognitive limitations. Examination of the Log File was required for these analyses. Owing to small sample sizes, statistical analysis could not always be used to show a level of significance of these results. Where such analysis was performed, the detailed results are shown in a table (with subsequent notes) in Appendix A: Statistical Analyses. In the remaining cases, the results are, by themselves, at least indicative, if not fully persuasive.

## A. RESULTS BY TEAM

Since several analyses performed on these data require balanced number of trials between the three independent variables, these analyses are based on the balanced set of 24 trials for each team, giving a total of 144 trials.

There is no *a priori* reason for supposing that the team ROC curve is symmetrical about the main, or negative diagonal (characteristic of discrimination between two equal variance normal (Gaussian) probability density functions (Swets, 1986:1)). Consequently, the best index of discrimination performance would be $A$, the area under the ROC (Swets, 1988), rather than $d'$, the distance from the minor diagonal to the ROC

along the main diagonal. For symmetrical ROCs, $d'$ and $A$ are directly related (Egan, 1975, p.81), so $d'$ may be used as a measure of discrimination performance. Calculation of $A$ generally requires use of a computer, given sufficient points to define the ROC. Obtaining several points on one ROC may be accomplished directly, by requiring the observer to adopt several different decision criteria, or, more efficiently, by use of a rating scale of, say, five levels (Swets 1986:2). This allows a single group of trials to be used to derive several points on the ROC, using the procedure outlined in Green and Swets (1988, pp. 99-103).

The confidence level in the CHIPS experiment provides six criteria (from high confidence neutral, through medium and low confidence neutral, low and medium confidence hostile, to high confidence hostile). However, the very small number of trials that results from dividing the data into six groups gives points that are not sufficiently accurate for calculating $A$; additionally, the confidence levels were not well used by the TAOs (based on personal observation), frequently being left at the lowest level simply to save time. Of the 189 trials where a final team (i.e., TAO) assessment was logged (time ran out before a final assessment was made in 3 trials), 116 had low confidence, 71 had medium confidence, and only 2 had high confidence. This would provide at best a four-point ROC.

A single point in ROC space does limit the ROCs which may pass through it, since a proper ROC is strictly non-decreasing (Egan, 1975, p.40). These bounds can be used to calculate the measure $A'$, which is also an appropriate measure of performance (see Norman, 1964). However, the calculation of $d'$ is much more straight-forward than $A$

or $A'$, and for ROCs that are only mildly asymmetrical, is still adequately indicative of performance.

1. **Comparison to Chance Performance**

Team performance was shown to be close to, but, in at least some cases, slightly better than could have been achieved purely by guessing. The number of correct initial and final correct assessments for the teams individually and together are shown in Table III. The corresponding $d'$'s for all teams together are also shown. The critical value of number of correct assessments is shown, based on the cumulative binomial distribution[14].

On the first assessment of each trial, the six teams together made the correct assessment in 82 of the 144 trials. This is significantly better than chance performance. By the end of the trial, performance was basically unchanged (certainly no better): the correct assessment was made in only 81 of the trials. This is also better than chance, but it is worth noting that the critical value for significance at p=0.1 is 80. The ROC also appears to be unchanged between initial and final assessments: $d'$ was a little under 0.35 in both cases.

At the team level, teams A, B, and F performed significantly better than chance on their initial assessments, but declined in performance to a level not

---

[14]The critical value shown in the table is the value of x which would give 1 - B(x;n=24,p=0.5) $\leq$ 0.1 for each team, or 1 - B(x;n=144;p=0.5) $\leq$ 0.1 for all teams, where B(x;n,p) is the cumulative binomial distribution, x is the number of "successes," n the number of trials, p the probability of "success," and 0.1 the significance level of the test ($\alpha$).

TABLE III: NUMBERS OF CORRECT ASSESSMENTS

| | | Team | | | | | | All teams |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | |
| Number of correct assessments | Initial | 15 | 17 | 11 | 13 | 11 | 15 | 82 |
| | Final | 13 | 13 | 11 | 16 | 15 | 13 | 81 |
| Critical value (p=0.1) | | 15 | 15 | 15 | 15 | 15 | 15 | 80 |
| | Initial | | | | | | | 0.349 |
| d ' | Final | | | | | | | 0.338 |

significantly better than chance by their final assessments. Team C remained constant between initial and final trials, at just under 50% correct. Teams D and E both improved significantly between initial and final trials (p=0.1; see Appendix A, p. 92), from chance level to a level significantly above chance (p=0.1).

It might be postulated that the very successful initial assessments for teams A, B, and F are the result of delaying longer (and thus obtaining more subordinate reports, with increased accuracy) before making the initial assessment. The TAOs were directed to log an assessment at least every 30 seconds, and so the initial assessment should have been made at the 30 second point, with absolutely no more than three probes available to the subordinates on which to base their reports. Table IV shows the mean time until the first TAO assessment, with standard deviations. Also shown in Table IV

are the mean number of probes that subordinates had made, and on which the reports made to the TAO were based, at the time of his first assessment[15].

TABLE IV: TIME AND PROBES AT FIRST ASSESSMENT, BY TEAM

| Team: | | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|
| Time of First TAO assessment | Mean | 39.56 | 46.30 | 40.19 | 43.03 | 42.09 | 58.78 |
| | Standard deviation | 7.67 | 6.80 | 5.44 | 5.25 | 7.22 | 5.63 |
| Number of Sub-ordinate Probes | Mean | 7.84 | 9.53 | 6.87 | 8.64 | 6.84 | 9.04 |
| | Standard deviation | 1.78 | 1.68 | 1.15 | 1.17 | 1.72 | 1.34 |

The times to first assessment did vary significantly between the teams ($p < 0.01$): team B was significantly ($p = 0.05$) greater than A and C, and F was significantly greater than all the other teams (see Appendix A, p. 94). The number of probes on which the initial assessment was based also varied significantly between the teams, with teams A, C, and E forming a low group, and B and F a high group. The TAO in team F consistently took longer for all his assessments, with an average assessment rate close to one per minute, where the other TAOs were closer to two assessments per minute.

Figure 4 shows the average time (beyond 20 seconds for clarity of presentation) of the first TAO assessment, the number of subordinate probes at first TAO assessment, and the number of correct initial TAO assessments (i.e., combining Tables

---

[15]Obtaining these values required examination of the Log Files.
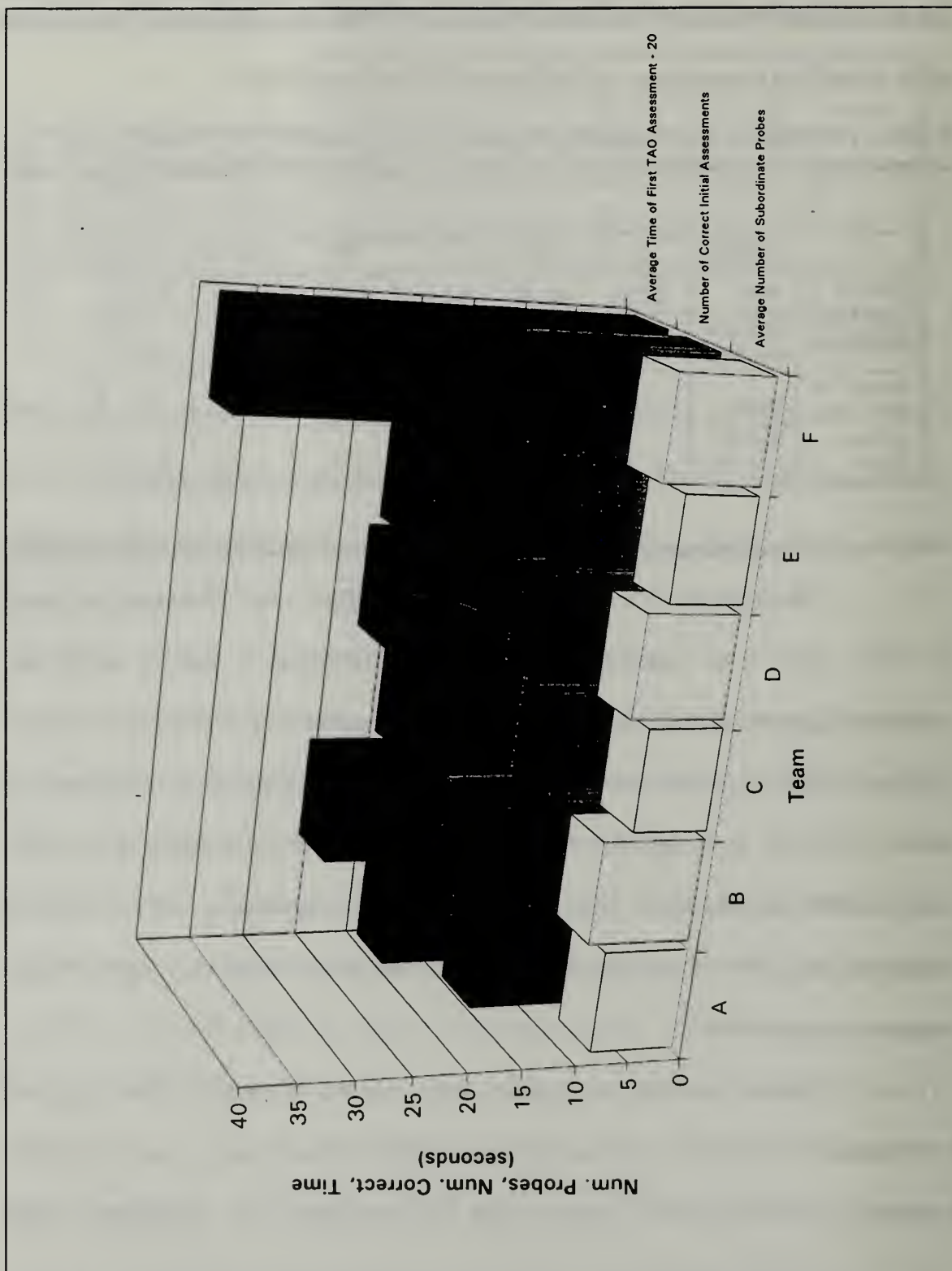
Figure 4: Comparison of Average Time to First TAO Assessment (minus 20 seconds), Number of Correct Initial Assessments, and Number of Subordinate Probes.

III and IV). This figure gives a strong impression of a relation between all the measures, but as detailed in Appendix A, p. 96, there is no significant relation to be found with time to the first probe. However, the number of subordinate probes is significantly related to the number of correct initial assessments ($r=0.88$, $p \leq 0.1$, see Appendix A, p. 99).

This result illustrates a general rule that will be seen again when considering the individual subordinate performance. There is a strong correlation between the time during a trial at which an assessment is made and the amount of information on which that assessment is based, as represented by the number of subordinate probes. For all the subordinates together, the correlation coefficient $r=0.95$, with 3602 degrees of freedom, which is significant at $p \ll 0.01$. However, measures of accuracy, in this case the number of correct initial assessments, are not in general related to the time in the trial, but are related to the number of probes available.

## 2. Effect of Independent Variables

The effects of the three independent variables individually on team performance are shown in Table V. The proportion of correct assessments is used as a measure of the accuracy of team performance. Details of the statistical analysis of significance of the differences are presented in Appendix A, p. 99.

When the TAO provides updates, performance degraded. The difference is significant only at $p=0.15$, so rejection of the null hypothesis of no effect by TAO update can not be confidently justified. A similar, small decline in performance when the TAO provides updates was also seen in Gough (1992). It is important to recall that

45

## TABLE V: AVERAGE PERFORMANCE VS. THREE INDEPENDENT VARIABLES

| | Proportion Correct | $d'$ | | Proportion Correct | $d'$ | | Proportion Correct | $d'$ |
|---|---|---|---|---|---|---|---|---|
| No Update | 0.597 | 0.50 | Low Risk | 0.611 | 0.57 | Low Stress | 0.687 | 1.12 |
| Update | 0.528 | 0.18 | High Risk | 0.514 | 0.09 | Med. Stress | 0.562 | 0.34 |
| | | | | | | High Stress | 0.437 | -0.36 |

this result is based on 24 balanced trials for each team, in all of which only two of the three subordinates had indications consistent with the actual hostility of the contact. Thus the subordinates were assessing indications that were conditionally independent. Higher risk also degraded performance[16], and this result was significant at p=0.1. As stress increased, performance again fell, also significant at p=0.1. These results are shown graphically (along with the change in $d'$) in Figure 5.

The question arises whether the decline in performance is caused by a change of the team ROC, or a shift in operating point on the same ROC. Use of an optimal decision strategy by the TAO would cause the team ROC to be determined not only by

---

[16]For risk, it could be argued that proportion of correct assessments is not the best measure, since the team is operating to minimize cost rather than maximize proportion of correct assessments. For comparison to the data on proportion of correct assessments, it should be noted that the average score in low risk trials was -0.389, while on high risk trials it was -0.931. For complete interpretation of the significance of this change, it would need to be compared to the expected value obtained by an optimal (normative) observer team, which is beyond the scope of this thesis.
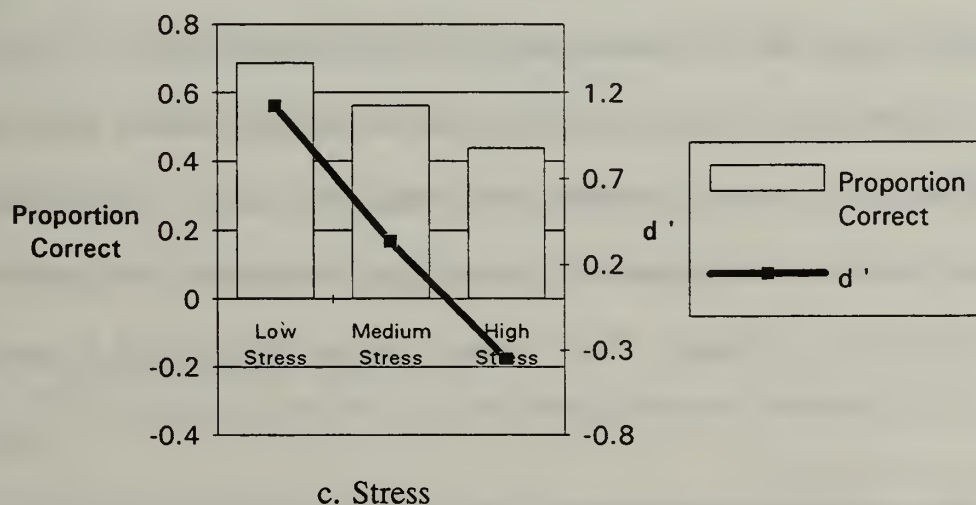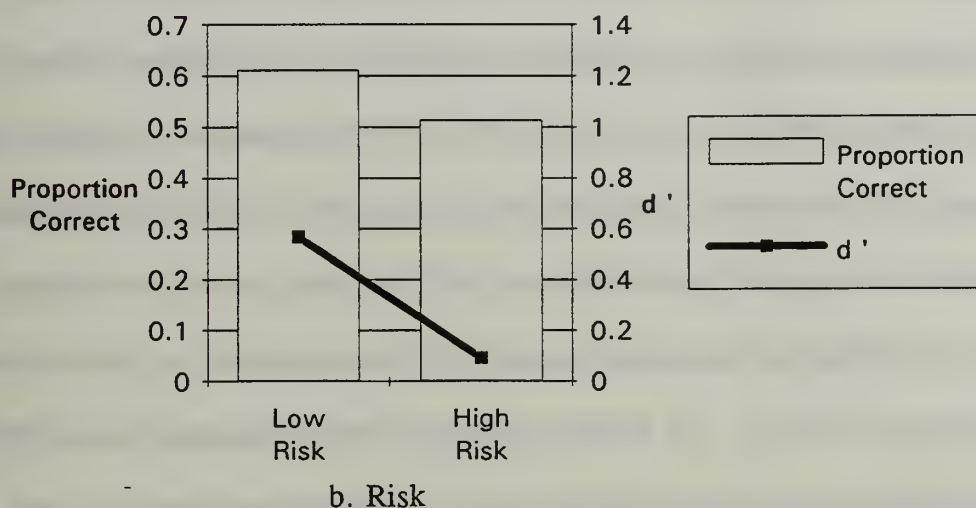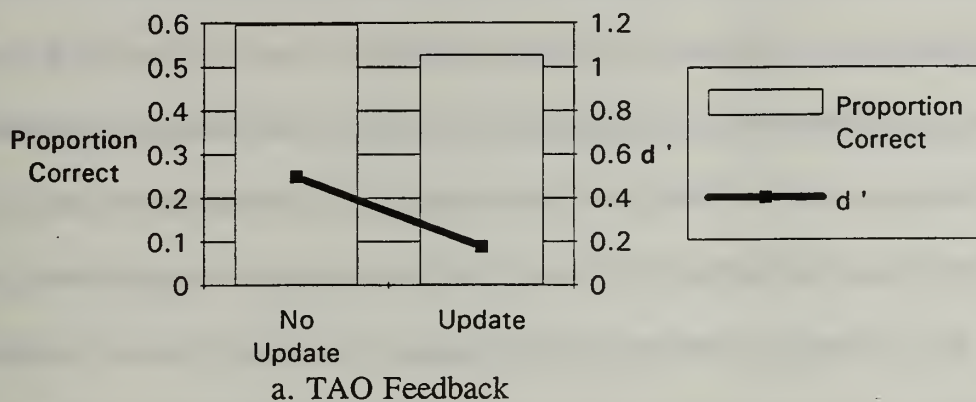
Figure 5: Effects of Independent Variables on Proportion of Correct Assessments and $d'$.

the individual subordinate ROCs, but also by the operating point on their ROC used by each subordinate (Pete, Pattipati, & Kleinman, 1993:2). However, the TAOs were observed to use a simple voting rule, regardless of reported confidence and subordinate behavior, throughout most of the trials. For example, if the subordinate reports were one indication of hostile with high confidence, and two indications of neutral with low confidence, the TAO would assess the target as neutral. A rigorous Bayesian analysis, with reasonable values assumed for high and low confidence,[17] would indicate that the target had a higher probability of being hostile (0.592 for the values given in note 17) than of being neutral. While it is true that in 24% of TAO assessments recorded by the computer the TAO chose a classification that was not consistent with the straight majority of subordinate assessments recorded at the time, the TAO assessment was very frequently based not on the subordinate assessment recorded by the computer, but rather on the assessments reported verbally. It is therefore assumed in this analysis that the team ROC was based only on the subordinate ROCs, not on operating points, so that a shift in operating point alone by the subordinates should not affect the team ROC.

TAO feedback should not change the subordinate ROCs, since the discriminability of the contact parameter being measured, which is conditionally independent of the other contact parameters, is unchanged. Consequently, the team ROC would be expected to be unchanged. The $d'$ measured for the team ROC did, however,

---

[17]E.g., a low confidence report of neutral based on the set of observations $x$ might indicate that $P(\text{neutral}|x)=0.6$, while a high confidence report of hostile might indicate that $P(\text{hostile}|x)=0.9$.

decline when the TAO provided updates. Since there is no reason to believe that the team ROC is symmetric, a change in $d'$ does not necessarily represent a change in ROC. Therefore, Figure 6 shows the No Update and Update points in ROC space, together with the corresponding boundaries for proper ROCs. The two points could, unquestionably, be from the same ROC if each fell within the boundaries for proper ROCs of the other[18]. This is not observed to be the case here; however, the variance of the points in space (and hence of the boundary lines) is not known. When the hit and false-alarm probabilities are calculated individually for each team[19] the standard deviation is found to be about 0.1 to 0.2, so it is possible that the two points are in fact from the same ROC, with some slight inaccuracy in their positions shown in ROC space. Indeed, even when an optimal decision-maker is modeled, and $d'$ measured based on 1000 assessments, the standard deviation in measurement of $d'$ is about 0.2, for n and sn distributions whose means are separated by $1.8\sigma_n$.

Like information structure (TAO feedback), risk would not be expected to change the subordinate ROCs or, in the case of an unweighted voting strategy, the team ROC, but would be expected to change the operating point on the ROCs. In high risk trials, however, the confidence of subordinates was observed to be used by TAOs on occasion, and reports of hostile parameters by subordinates were weighted more heavily

---

[18]This technique is derived from the ranking procedure described by Norman (1964).

[19]Note that this calculation does not give a well refined result: there are only six hostile contacts for each team for each condition of TAO feedback. Thus there are only seven possible values for p(Hit|sn) and for p(False-alarm|n).
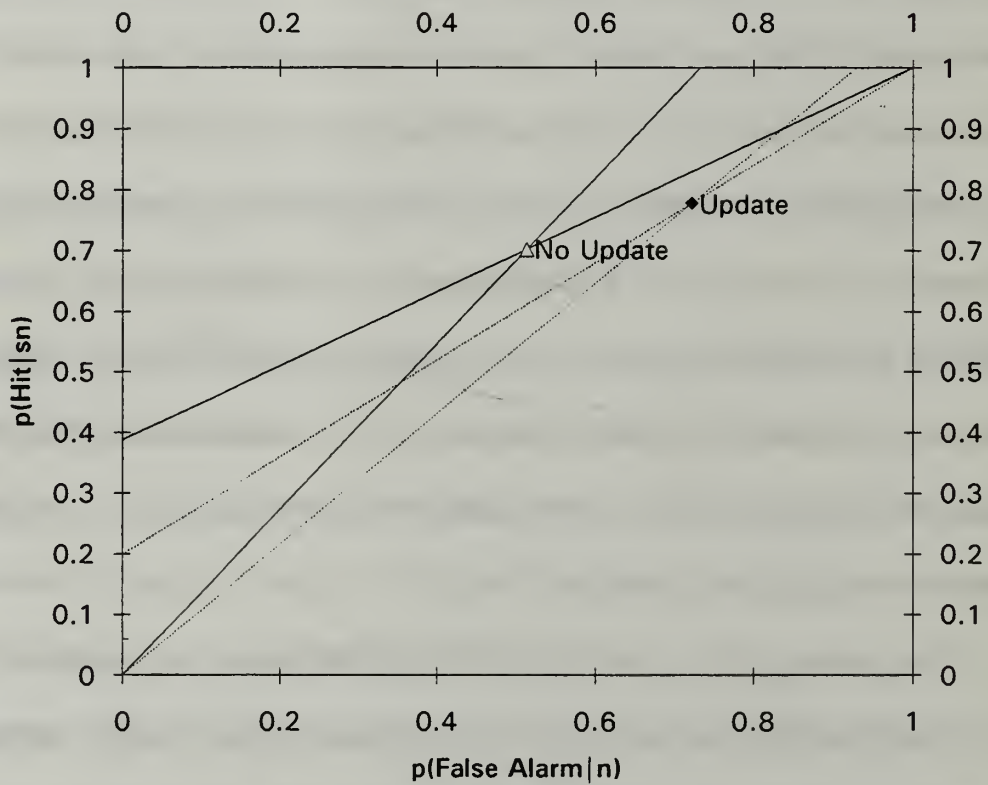
Figure 6: Comparison of ROC Boundaries for No Update and Update Trials.

than reports of neutral parameters[20]. Therefore the shape of the team ROC could alter, without a change in the shapes of the subordinate ROCs. Certainly, the subordinate operating point on their ROCs would be expected to change, since the expected cost of a miss has changed. In the end, the only conclusion that can be readily drawn from this figure, as expected by the calculated value of $d'$, is that performance in high risk trials is very close to chance level.

The change in subordinate ROCs with increasing stress corresponds with the next section of this chapter, and so is not considered further here. The decline in $d'$ with increasing stress should be attributable to the decrease in information available to the subordinates, and thus to the accuracy of information reported to the TAO. Note that there can be no proper ROCs passing through the high stress point, since this represents performance below the chance level. Performance could have been improved in this case simply by reversing each decision!

## B. RESULTS FOR INDIVIDUAL SUBORDINATE ROLES

During the conduct of the CHIPS experiment, it was noted that the TAO for team F consistently made his final decision early in the trial, often with 30 seconds to one minute remaining before the end of the trial. By contrast, most other TAOs waited as

---

[20]These observations are based purely on personal observation of all the TAOs. From the data recorded by the computer, only 43% of the assessments that were not a straight majority rule occurred during high risk trials. Again, it should be stressed that the information available to the TAO recorded by the computer is not necessarily representative of the information used by the TAO to make his assessment, because of the large amount of verbal reporting performed, particularly just prior to a TAO assessment.

long as they could, to gather as many reports from their subordinates as possible, before making a final decision. Team F's performance was not outstanding (see Table III), but it appeared to the observer that subordinate assessments were more accurate than had been seen with other teams. Indeed, it had appeared that as low stress (three minute) trials neared completion, subordinate confidence, which had been gradually growing, started to decline, with frequent reports of low confidence. This lead to the postulation of the theory that subordinate performance increased with time only up to about 2 minutes into a trial, which corresponds to six or seven probes, at the average subordinate probe rate of 3.14 probes/minute. Were we seeing another manifestation of the "Magical Number Seven," which so plagued Miller (1956)?

Extraction of the data from the Event Log Files was required in order to analyze each individual probe. When the Log File from all the trials are combined, 871 pages of data are available, of which most is irrelevant. A BASIC language program was written to extract the important data from these pages, reducing the information to 8,571 events -- either probe or assessment. In addition, since the number of probes available to the subject at time of assessment was not recorded by the computer, the BASIC program was designed to report the number of probes that had been made by the same player during the trial, prior to ten seconds before the assessment[21]. Since this analysis

---

[21]This count overestimates the actual number of probes available for assessments early in the trial. The subjects used the technique of initiating a probe (with its associated ten second delay) just before assessing the results of the previous probe. This allowed more frequent probes, because the ten second delay ran concurrently with the time spent assessing previous probes. However, if more than ten seconds was spent before logging the assessment, the probe information would be "available,"

examines only the subordinates, and does not require balanced trials between the independent variable (indeed the independent variables are ignored as much as possible), all 32 trials for each team were used.

The first aspect of these data to be examined was how the proportion of correct assessments varied with time into the trial. Low stress trials were used to give the longest time period for examination. Time of assessment is recorded to the nearest second, so the proportion of all assessments recorded for each second that were correct was calculated. This gives a coarse graph, since there may be very few probes in any given second, so the data were smoothed. The same smoothing method, a running average, was used throughout this analysis, for simplicity.[22] Each value was replaced by the average of itself and the two adjacent values. This process was repeated as few times as necessary to give a meaningful plot.

When this analysis was performed (using seven smoothing iterations), the result was disappointing: see Figure 7a. An initial improvement in performance during the first minute was followed by widely varying performance that was, nevertheless, approximately constant with time.

However, as was shown in the previous section, time is not a good analogue for amount of information available, despite the high correlation observed. Therefore, the

---

and counted by the BASIC program, even though it would not be shown to the subject during the assessment in question.

[22]A better (but more complicated and time-consuming) technique might have been to perform a Fourier transform of the data, remove high-frequency components, and then perform an Inverse Fourier transform.
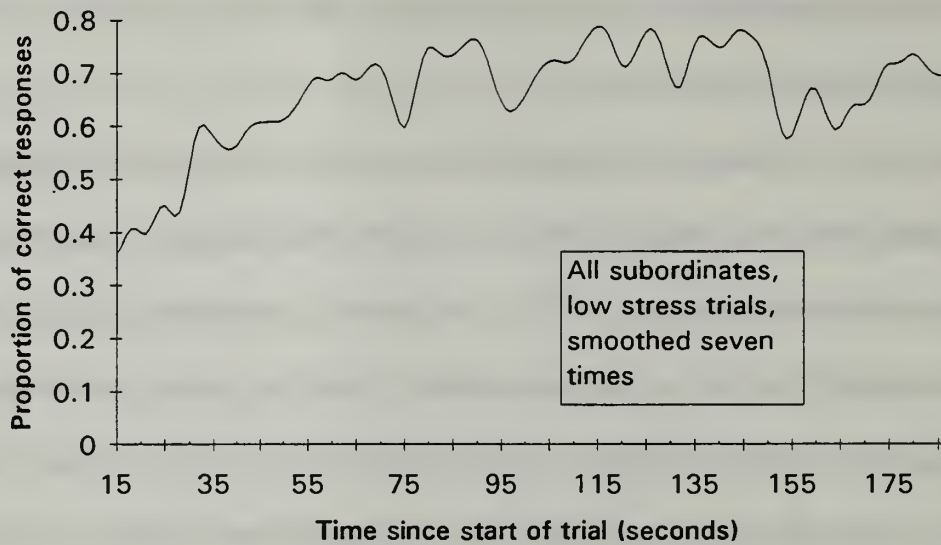
a. Accuracy as a Function of Time
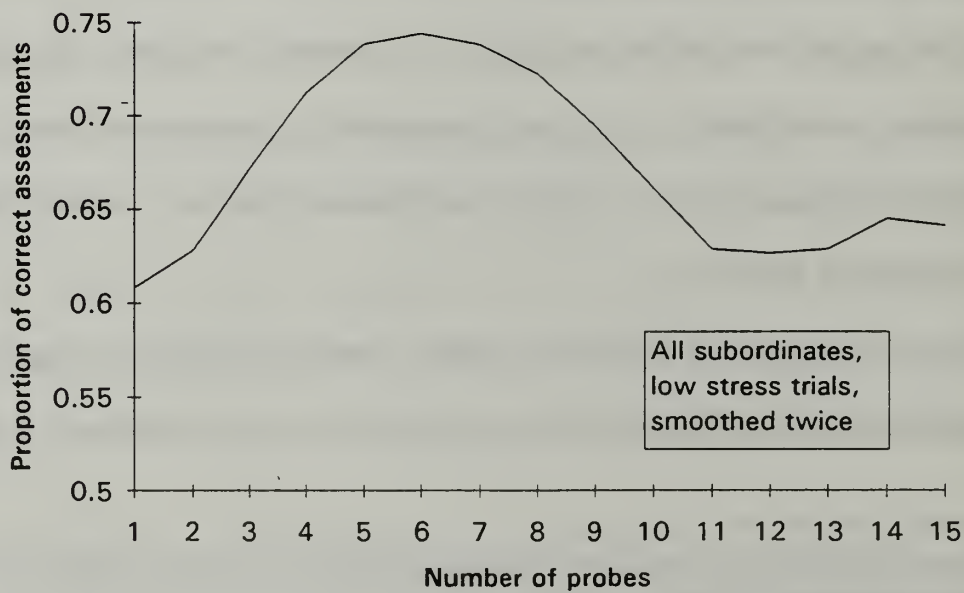


b. Accuracy as a Function of Number of Probes

Figure 7: Accuracy as a Function of Time and Number of Probes

same analysis was performed using the number of probes available when the assessment was performed, rather than the time since the start of the trial (see Figure 7b). This produced (with only two smoothing iterations) a graph very much as expected, with a peak in performance at five to six probes.

Recalling the original, informal, observation that confidence reaches a peak and then declines, the average confidence of subordinate assessments is plotted against number of probes in Figure 8. Also shown is the number of assessments that were made based on the given number of probes (for all stress levels). Average confidences are shown for both low stress trials and all trials, demonstrating the variation with stress level is minor. Confidence, like accuracy, is seen to peak and then decline, in general, as the trial progresses. The peak in confidence occurs slightly later than in performance.
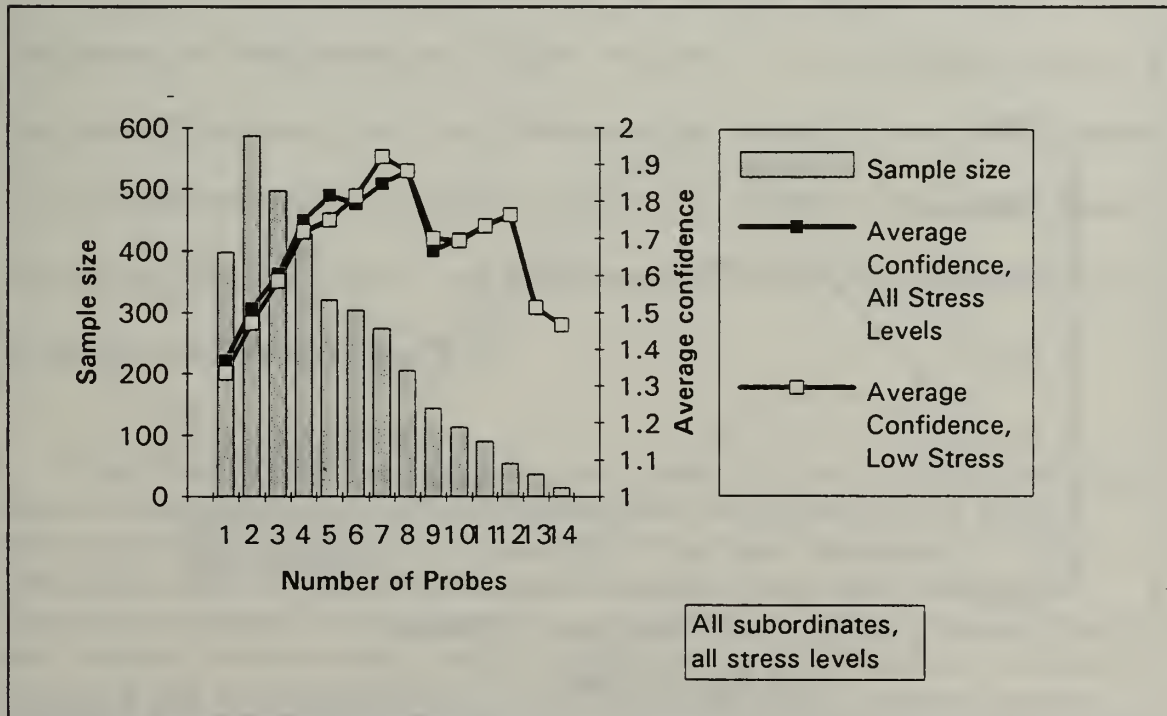
Figure 8: Average Confidence as a Function of Number of Probes.

As can be seen in Figure 9, the degree to which this effect is seen varies with the subordinate role. The remainder of this chapter will examine the individual subordinate roles in detail, looking in particular at how the ease with which hostile and neutral targets can be distinguished (as measured by $d'$) varies with amount of information available (as measured by number of probes).

### 1. Identification Supervisor (IDS)

#### a. *Proportion of Correct Assessments*

Figure 9 shows that there is very little variation in the proportion of correct assessments with number of probes for the IDS during low stress trials. Further analysis will combine stress levels, so that the data may be separated by team yet retain
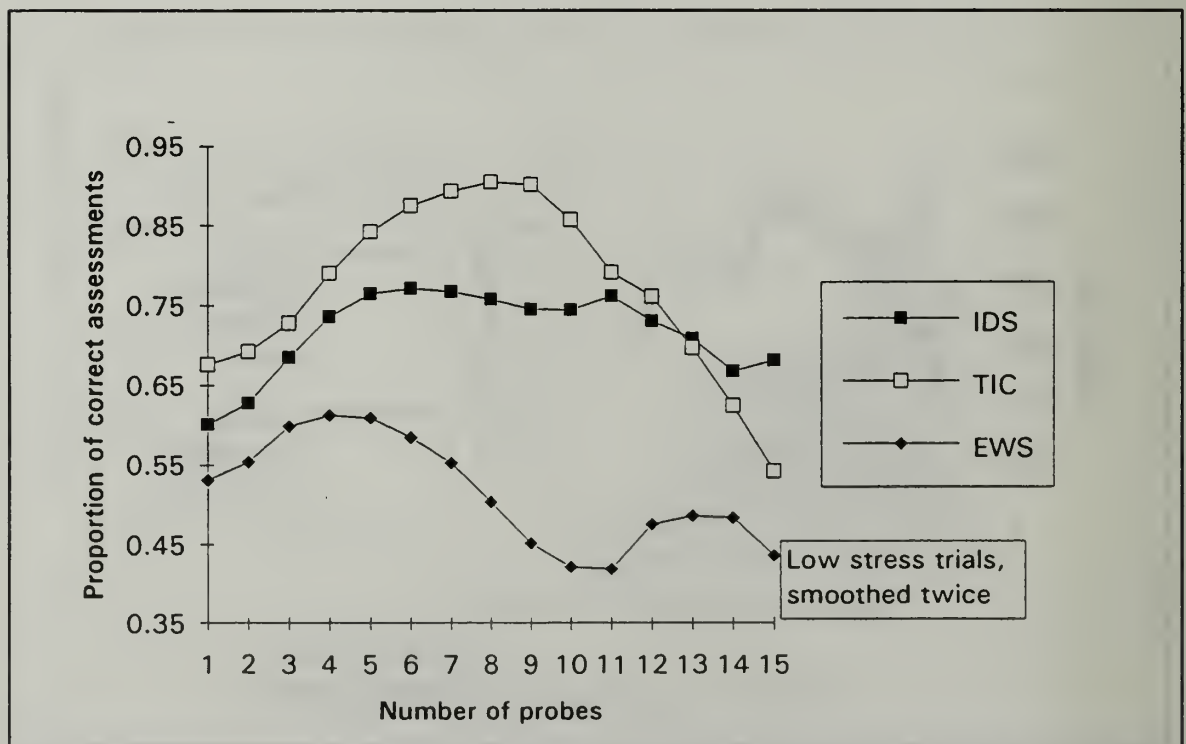


Figure 9: Variation of Accuracy with Number of Probes--Subordinates Individually.

sufficient data for calculation of $d'$. Figure 10 shows the results for each stress level, and the average. As can be seen from the figure, and from the analysis shown in Appendix A, p. 101, there are significant differences between the low, medium and high stress results (e.g., p=0.078 between high and medium stress, p<0.01 between low and medium stress). The same analyses do not show any variation with number of probes, and the plot of average proportion of correct assessments is relatively flat. Despite these significant differences with stress, the data were boldly combined in subsequent analysis.

### b.  *Ideal Observer Performance*

The task of the IDS follows the form of SDT very closely. This task consists of distinguishing between two normal, equal variance distributions: one with a mean value of 40 (a hostile target in CHIPS, n (noise) in SDT) and one with a mean value of 60 (a neutral target in CHIPS, sn (signal plus noise) in SDT)[23]. Since the distributions are both normal, and of equal variance, $d'$ is an appropriate measure of discriminability between them. One of the advantages of using $d'$ is that, while changes in the independent variables (in particular risk) may be expected to change the operating point on the subordinate's ROC, and hence the proportion of correct assessments, they should not affect the underlying ROC.

---

[23]Traditionally, SDT uses a positive signal, so that the distribution with the higher mean is the sn distribution, and this convention has been preserved here. This arrangement would reverse the interpretations of "hit" and "false-alarm" in CHIPS. The solution used is to preserve their correct meanings (e.g., "hit" is the correct detection of a hostile), and manipulate the calculation of $d'$. The arrangement itself is less important than the concept, merely requiring care in the calculations: the check to make is that, if the probability of a hit is higher than that of a false-alarm (as is desirable), then $d'$ should be positive.
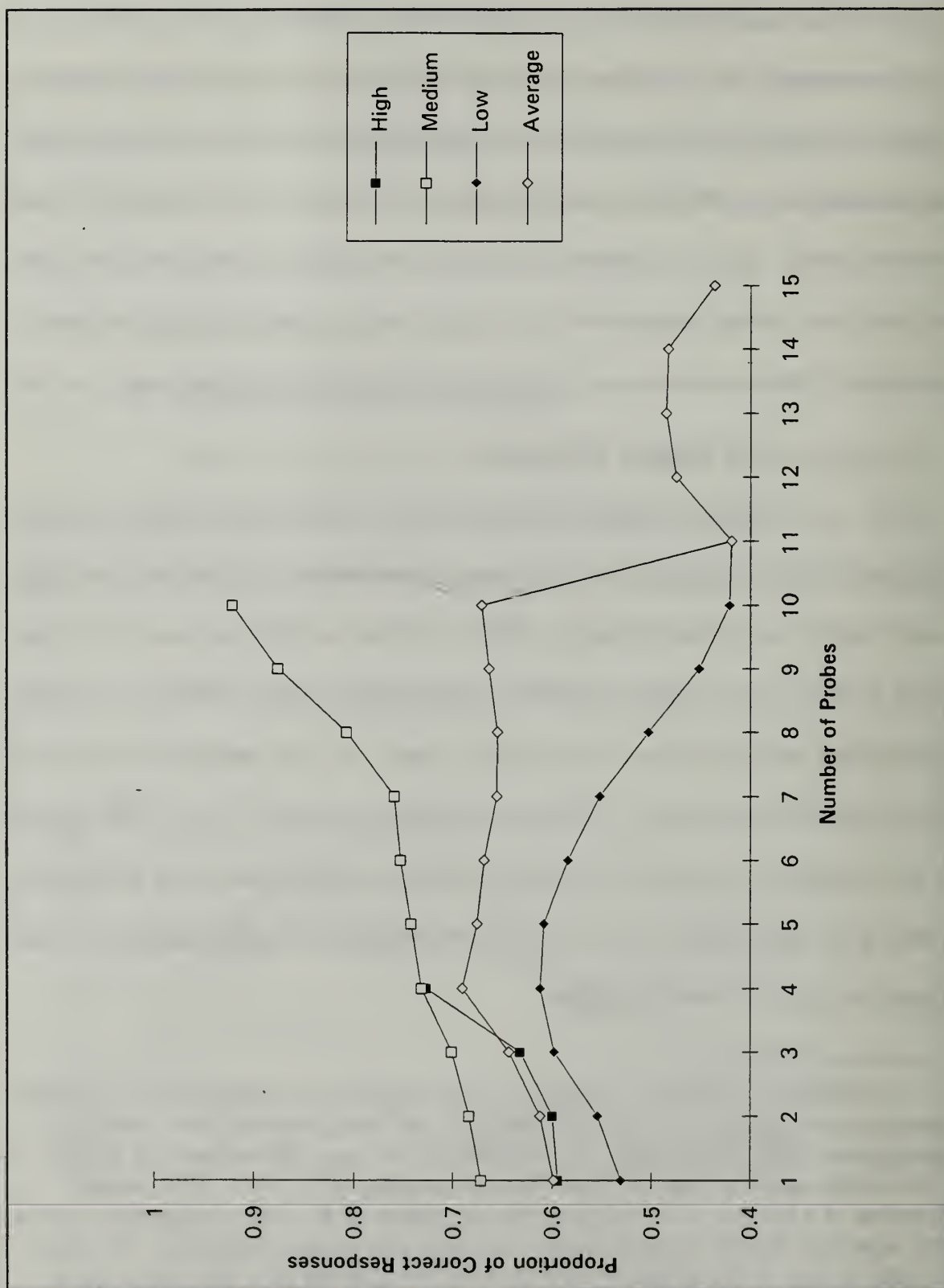
Figure 10: Proportion of Correct Responses for IDS, by Level of Stress.

The real size of the target is a random variable, constant throughout each trial. It is derived from one of two distributions: small sizes from a normal distribution with mean 40 and standard deviation 3; large sizes from a normal distribution with mean 60 and the same standard deviation. Thus even an ideal observer with access to the real size of the target will not have perfect discriminability: the maximum achievable $d'$ is:

$$d'_{max} = \frac{\mu_{sn} - \mu_n}{\sigma_n}$$
$$= \frac{60 - 40}{3}$$
$$= 6\frac{2}{3}$$

(IV-1)

Each time the size is probed by the subject, the value displayed is further corrupted by noise, which is a random value drawn from a normal distribution with mean 0, and standard deviation 20. The ideal observer, then, for each individual probe is distinguishing between normal distributions with a standard deviation of $\sqrt{(20^2 + 3^2)} = \sqrt{(409)} \approx 20.2$. After $N$ probes, the average of the observed values will have a smaller standard deviation:

$$\sigma_n(N) = \sqrt{3^2 + \frac{20^2}{N}}$$

(IV-2)

Therefore, the $d'$ demonstrated by the ideal observer will increase with the number of observations, up to the maximum value shown in equation (IV-1), as:

$$d'(N) = \frac{60 - 40}{\sqrt{3^2 + \frac{20^2}{N}}}$$

(IV-3)

This situation corresponds to the presence of random noise, and constant noise (i.e., the variability of actual target size), for which the improvement in $d'$ with $N$ is observed to be less than with random noise in human subjects (Swets, et al., 1959). The result (see Figure 11) is ideal performance that is not quite proportional to $\sqrt{N}$, as it would have been with only random noise present; however, the departure from linearity is sufficiently slight to be largely irrelevant[24].
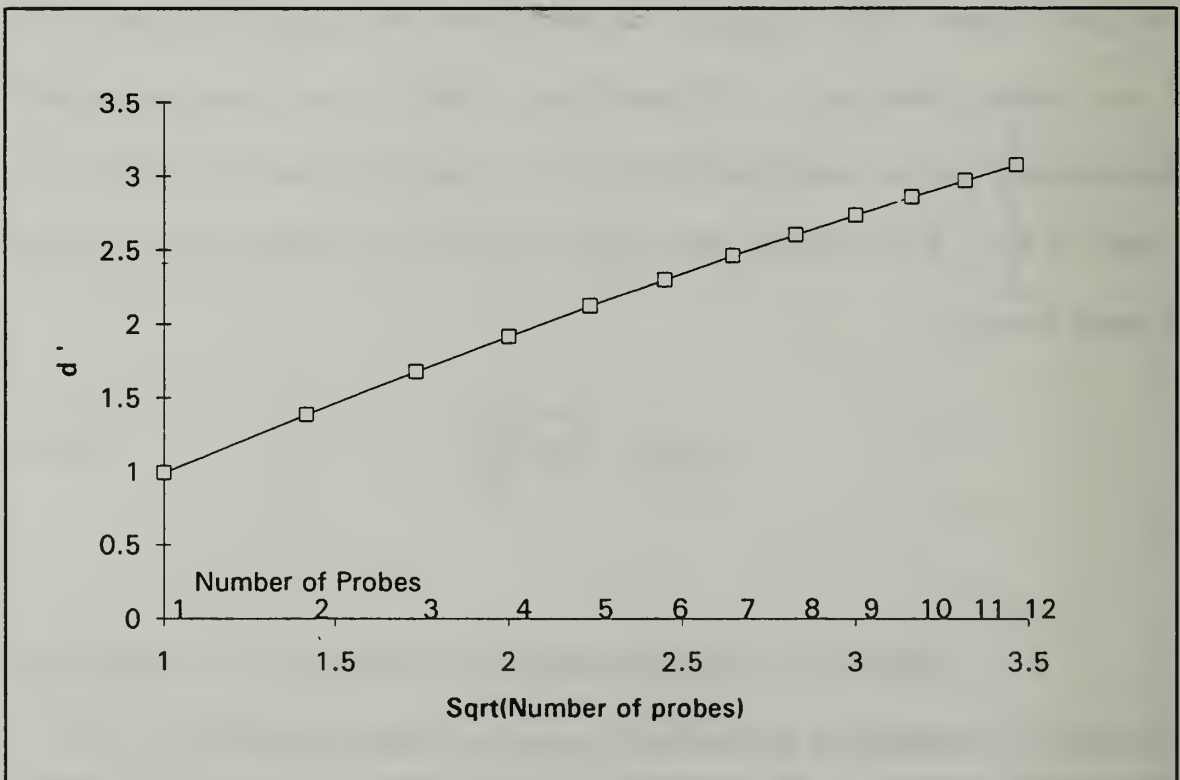


Figure 11: Ideal IDS Observer Performance

_____

[24]This is shown later in Figure 15 on p. 64.

### c. Selection of Averaging Technique

The ideal situation would be to calculate a $d'$ for each subject for each condition of the independent variables. There were exactly one hostile and one neutral target (sometimes plus one or two more targets either hostile or neutral, from the set of eight "distractors") presented to each subject in each condition of the independent variables. Consequently there were insufficient data to estimate a probability of hit and probability of false-alarm at each probe for each condition of the independent variables. Team E, for example, did not have <u>any</u> small, low-stress targets. Even disregarding the independent variables, as was done, problems with low data counts, or very successful subjects, were encountered. Not infrequently, even when a significant number of data existed, there were either perfect hit rates or false-alarm rates (1.0 or 0.0, respectively), which could not be transformed to Z-scores (which is required for calculation of $d'$). The hit or false-alarm rate was replaced with 0.99 or 0.01 respectively in these cases, whenever there were five or more opportunities for a hit or false-alarm.

Grouping of the number of probes in pairs was used to raise the accuracy of calculation of hit and false-alarm rates, for some analyses (see Figure 12, for example). However, generally $d'$ could be calculated at each probe.

There are two ways to average data across teams--which is the same as averaging data across observers-- shown in Figure 12. Either the $d'$'s could be calculated individually, and then averaged directly, or the hit and false-alarm rates could be calculated for all cases regardless of team, and these combined rates used to calculate the "collapsed" $d'$. The latter technique calculates a more accurate $d'$, since it is based on

61

more data, and is less likely to require estimation for perfect hit or false-alarm rates. However, as Macmillan and Kaplan (1985) explain, it leads to a lower estimation of $d'$, and the amount of underestimation increases with the difference between the operating points of the subjects on their ROC. Figure 13 shows the individual operating points for the six teams in ROC space at one, two, three, and four probes, along with the ROC of the optimal observer. As can be seen, all observers were relatively close to the ideal ROC, but appear to be operating with widely different criteria. Consequently, the collapsed $d'$ calculates a much reduced value, that does not track well with the ideal observer (see Figure 12). The average of the individual team $d'$'s will therefore be used in this study.
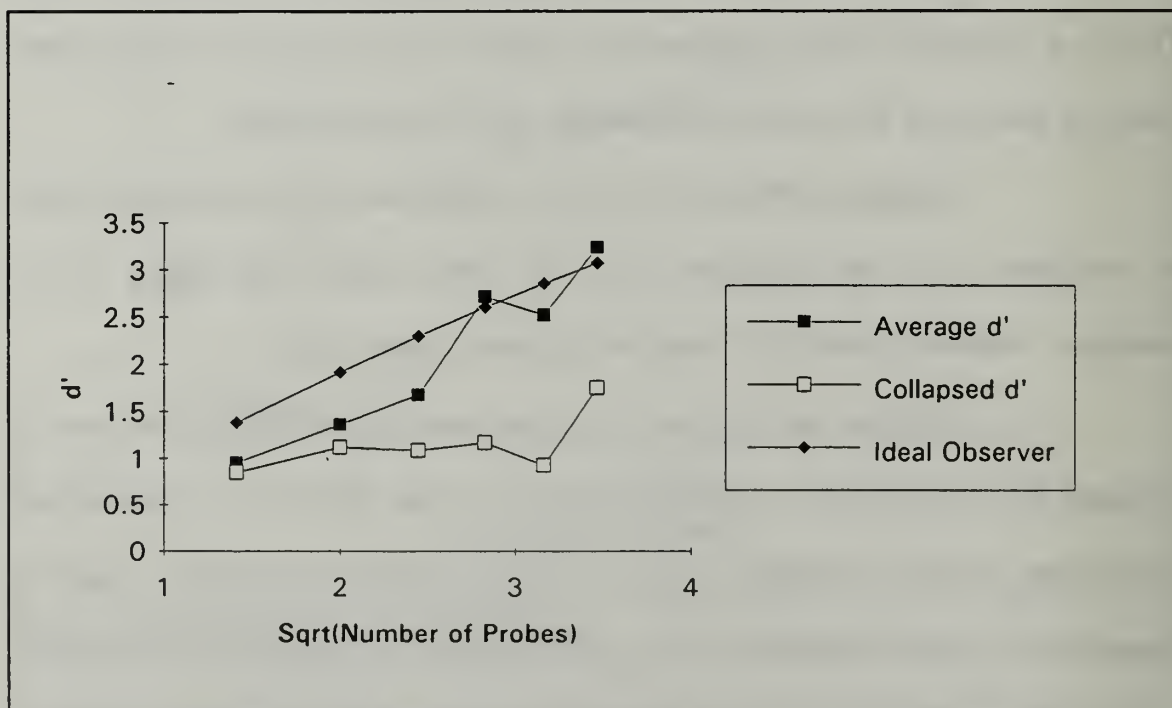


Figure 12: Comparison of Average and Collapsed $d'$.
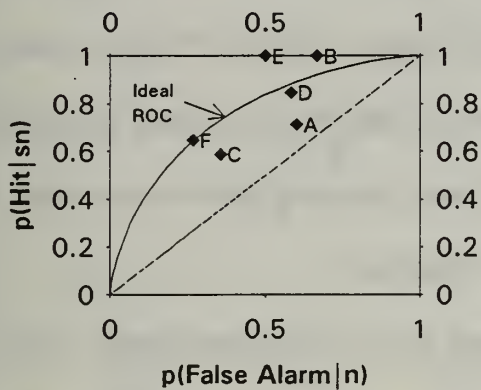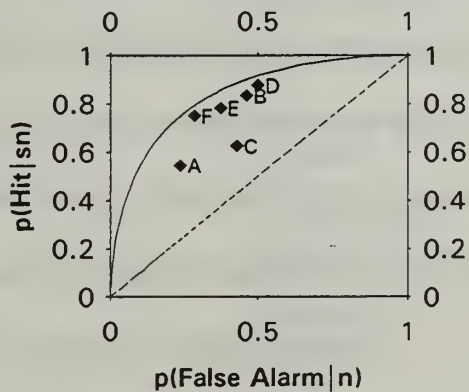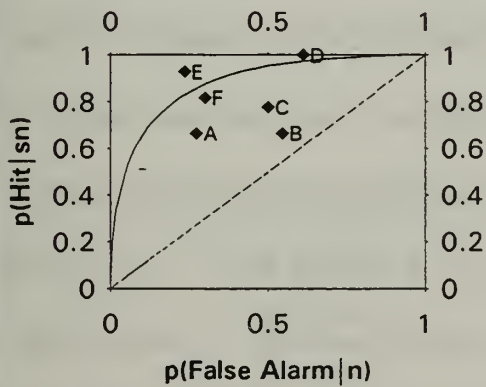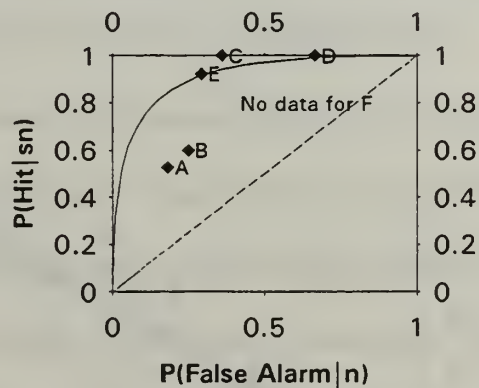
Figure 13: Operating Points in ROC Space for Each Team--First Four Probes

### d. Results for IDS

The performances of each individual team, by pairs of probes, is shown in Figure 14, along with the performance of an "ideal observer." As can be seen, initially performance is somewhat below ideal, and variable. By the 12th probe, there is less variability (and also fewer teams, since for four of the teams decisions were made almost always before the tenth probe, which would account for the reduction of variability) and performance tracks more closely with ideal.

The averaged $d'$ is shown in Figure 15, for individual probes. Again, ideal performance is shown, with the regression line for the ideal performer plotted.
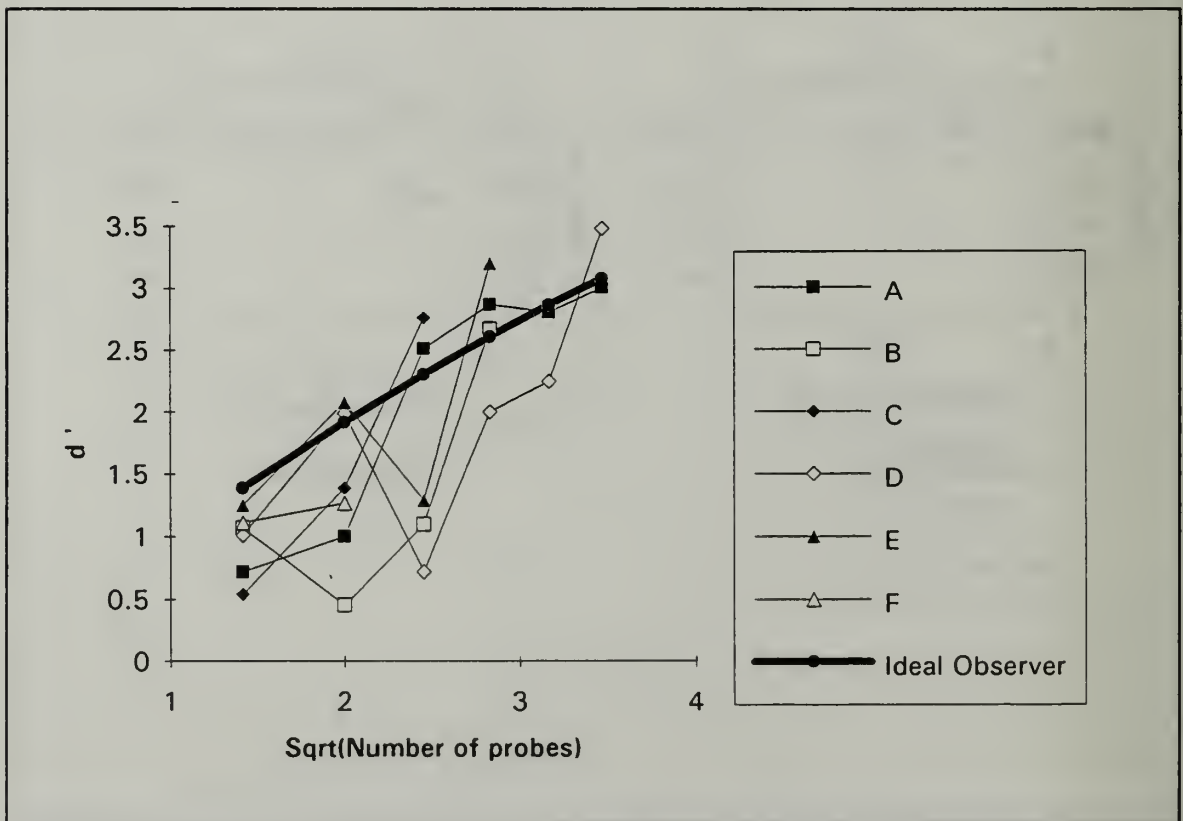


Figure 14: Performance of Each IDS vs. Number of Probes
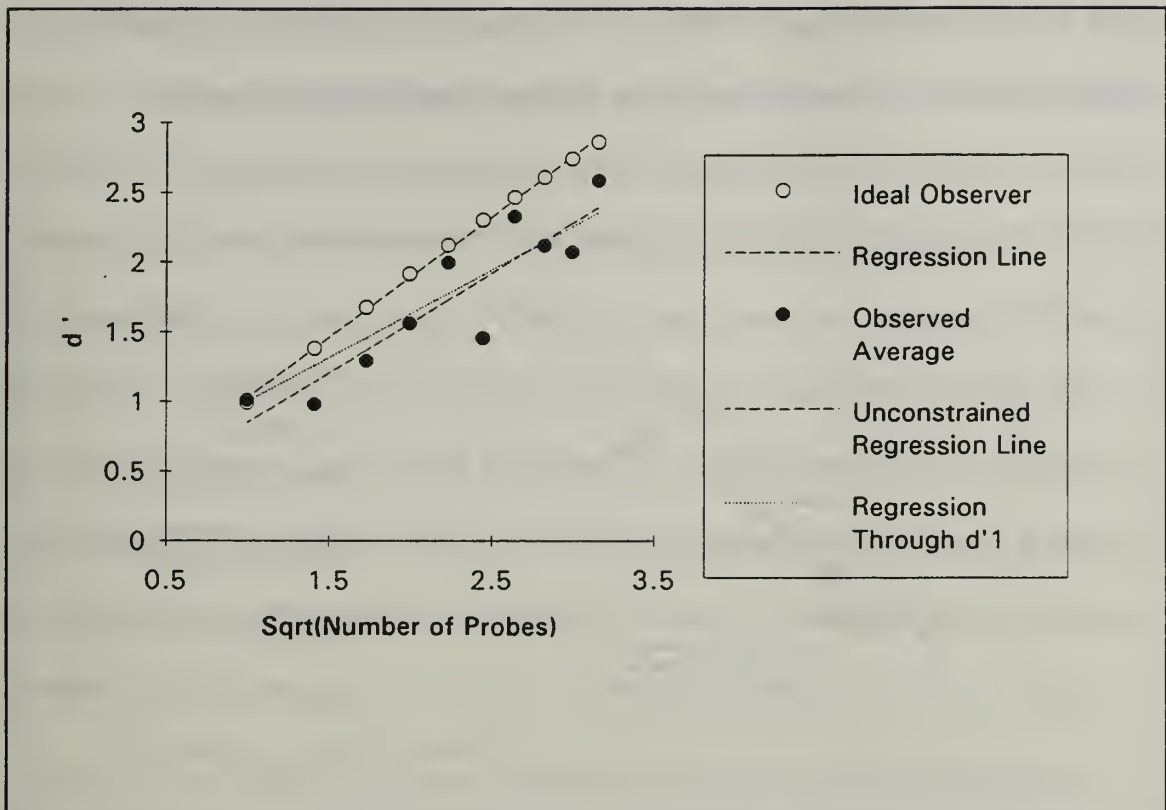
Figure 15: Average IDS Performance, with Two Alternative Regression Lines.

This figure shows how closely ideal performance may be approximated by a straight line.

Two regression lines for the observed data are shown. One is the best fit, unconstrained

regression line. As can be seen, it is very close to being parallel with the ideal line, but

is slightly below it. This would indicate a subject who improves at the same rate as the

ideal observer with increasing numbers of probes, but is uniformly less able to

distinguish between the distributions.

Mathematically, a uniformly lowered $d'$ would imply that the means of

the noise and signal (neutral and hostile) distributions were closer together for the actual

observer than for the ideal observer (see Figure 16). Since the means of the distributions

were well defined (40 and 60), and there is no possibility that the signal perceived by the

subject differed from the signal presented, as is postulated in psychophysiological Signal Detection Theory, there would have to be a different mechanism at work here.
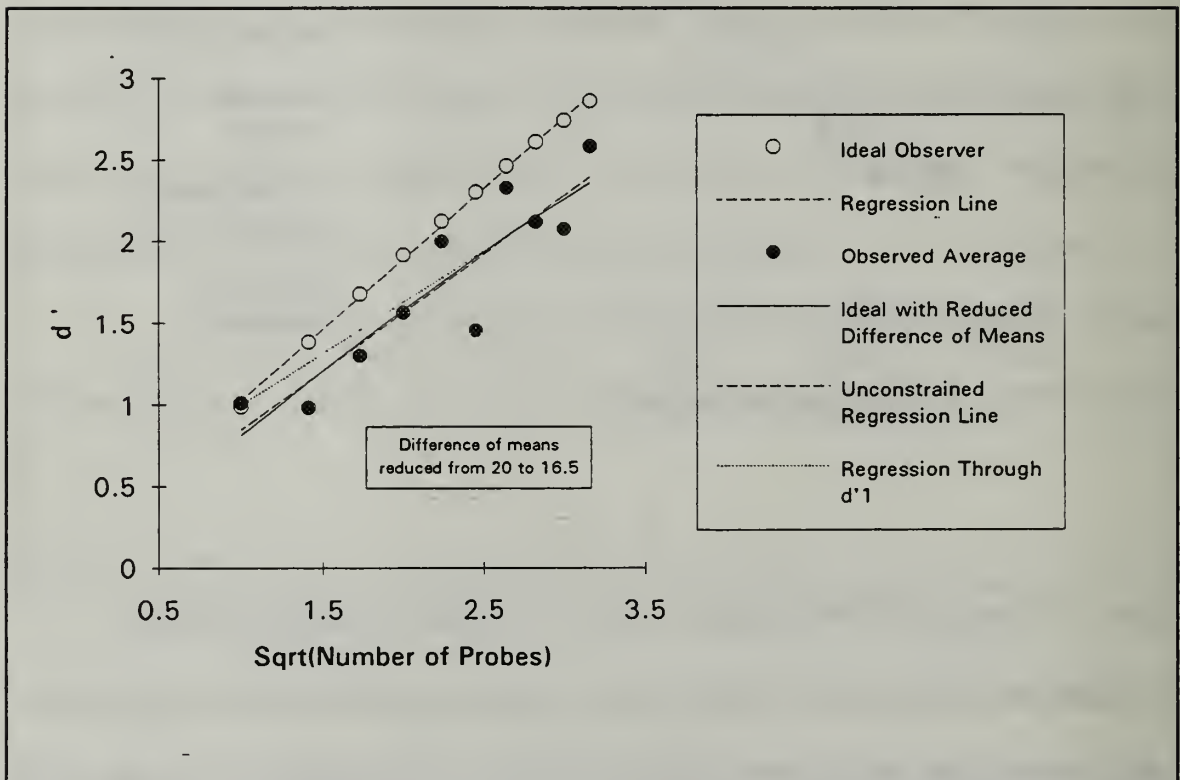


Figure 16: Average IDS Performance, Showing Reduced Difference of Means Model.

The closeness of the observed initial $d'$ to the ideal is compelling[25], and leads to the second regression line shown in Figure 15. This line is constrained to pass through the ideal initial $d'$. Like the first line, the fit is reasonably good, but now the slope is significantly ($p < 0.1$) different from the ideal performance (see Appendix A, p. 104). There are many models that would fit this result. An observer able to make each

---

[25]Especially since the initial observed value should be computed nearly accurately, and need only be compared consistently to a criterion value (e.g., and optimally for low risk trials, 50) to achieve the optimal $d'$.

observation well, but unable to integrate observations optimally, would perform in this manner. Also, an observer who introduced extra, internal, noise would show similar behavior. This explanation is compelling, since the requirement to multiply two numbers in order to arrive at the observation would introduce arithmetic inaccuracies, that could be represented as internal noise. When the ideal observer is further hindered by introducing "arithmetic" noise with a mean of 0, and a standard deviation of 5, the resulting performance is as shown in Figure 17. This is very close to the observed performance. It is also very unlikely that observers were able to maintain an accurate running average of observations. Further discussion of strategies is to be found in Chapter V.
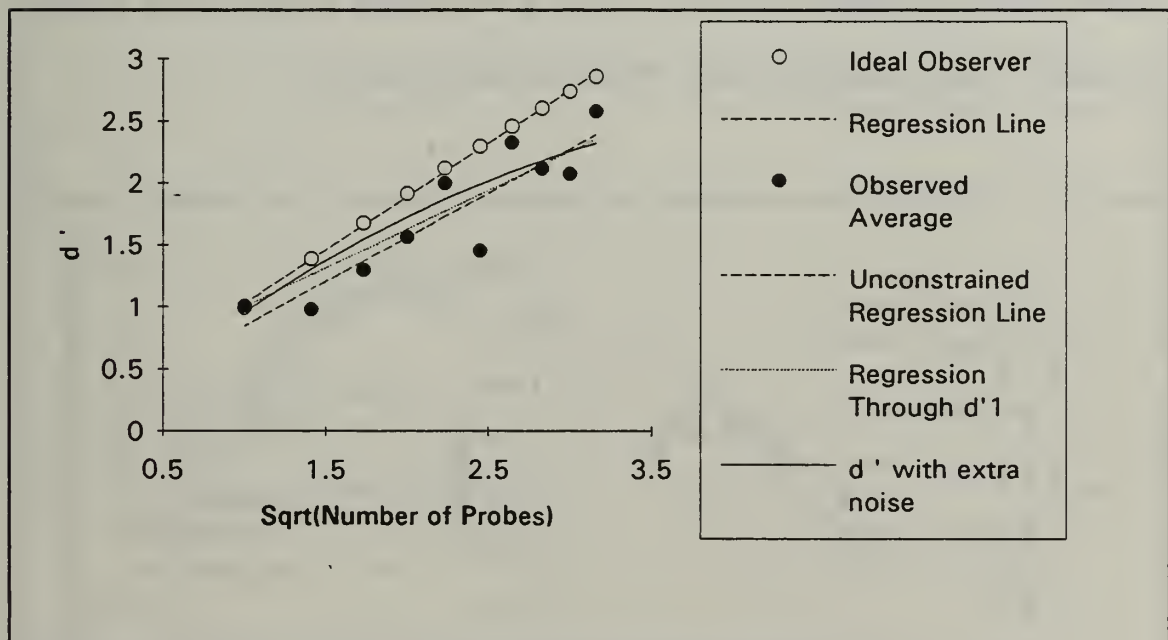


Figure 17: Average IDS Performance, with Extra Noise Model.

An inability to average more than five to nine numbers, as might be implied by short-term memory limitations (or simply by the fact that only the five most

recent probe results were displayed to the subject by the computer) would result in performance that leveled after five to nine probes. While it is difficult to be certain, with at most ten probes by any subject, the evidence for this limitation is not compelling in the case of the IDS.

One of the indications that led to tracking $d'$ against the number of probes was the decline of confidence towards the end of (low stress) trials. Therefore, it is interesting to compare $d'$ against the reported confidence. The results are shown in Figure 18. Close agreement can be seen throughout, except for the very last probe. The correlation coefficient between $d'$ and confidence is 0.679 which, with eight degrees of freedom, is significant at $p < 0.05$. This may indicate that the success actually seen on the last probe is an anomaly, since it is based on the smallest amount of data of any of the probes. This would significantly change the analysis above.
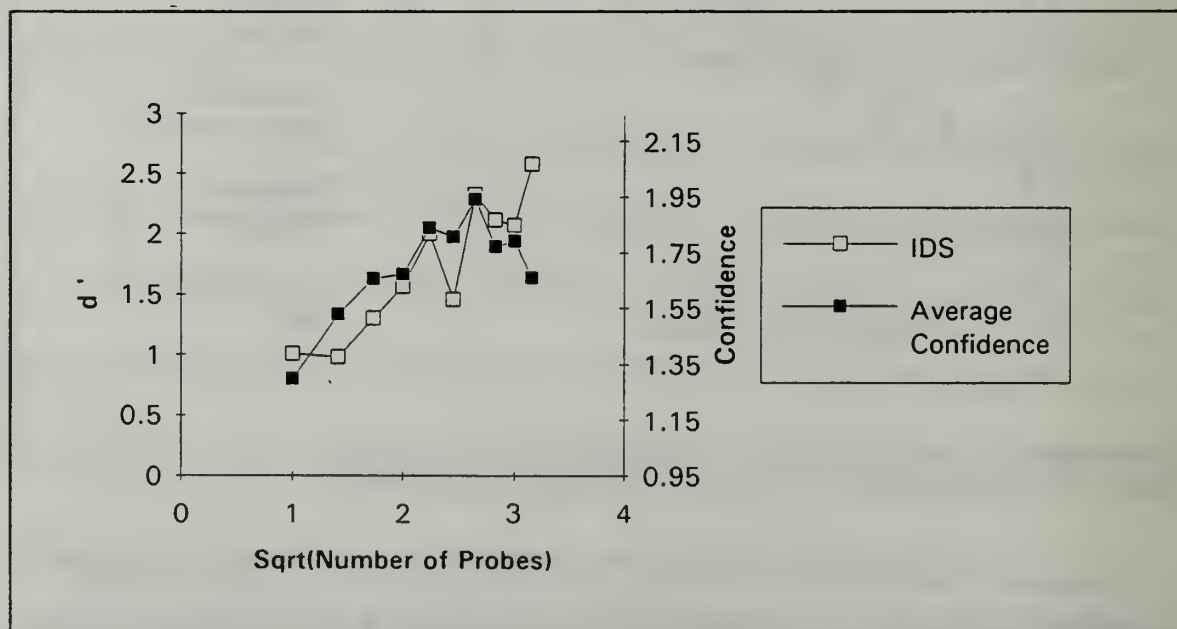


Figure 18: IDS $d'$ Compared to Average Confidence.

## 2.  Target Identification Coordinator (TIC)

### a.  *Proportion of Correct Assessments*

The TIC showed the greatest tendency to peak and then decline in performance of any of the three subordinates, as seen in Figure 9. Furthermore, Figure 19 shows that there is very little difference in the proportion of correct assessments between the different stress levels. Of course, the entire decline past 10 probes is based on very few observations, of only one subject, and cannot be investigated in terms of $d'$, because there are insufficient data to estimate hit and false-alarm rates accurately.
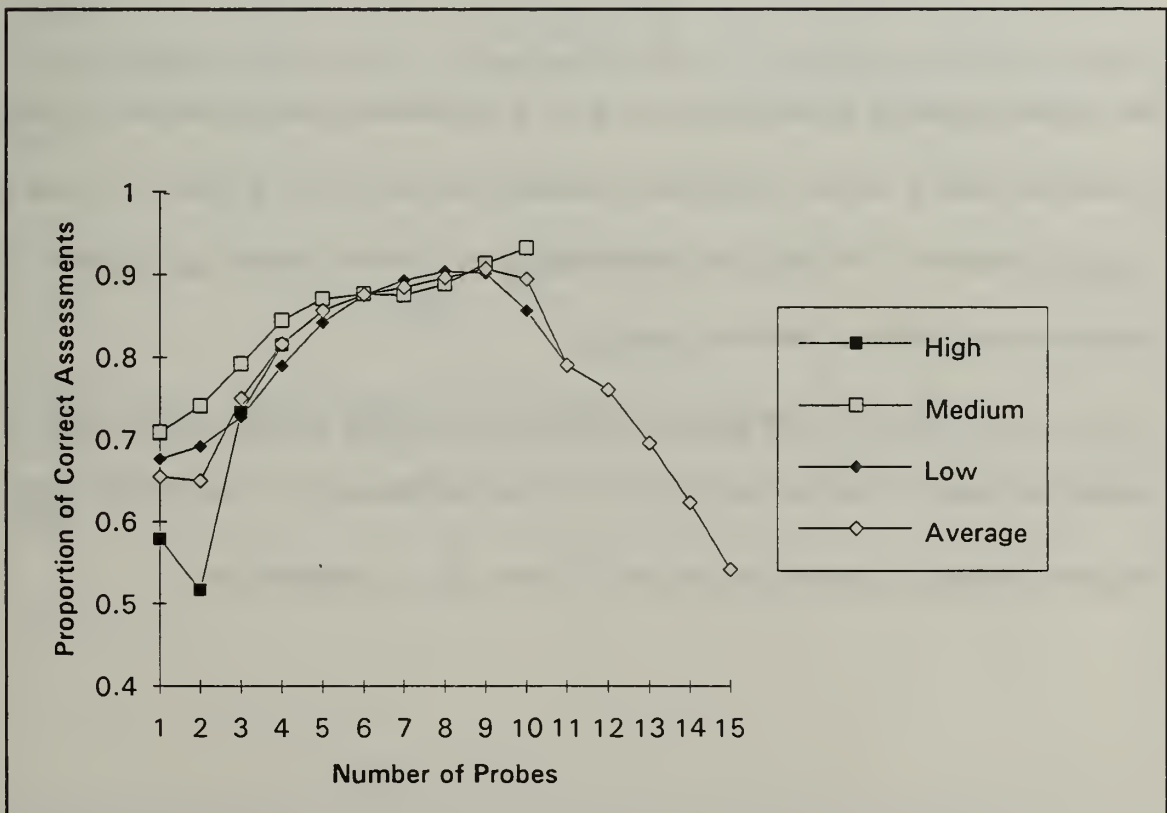


Figure 19: Proportion of Correct Assessments for TIC, by Level of Stress.

### b.   *Ideal Observer Performance*

The normative model for the TIC is the most complex of the three subordinates, and there exist two "optimal" strategies that could be used, one of which is far more efficient than the other.

Each time the target is probed, a descent rate (random variable $D_i$) is sampled from one of two normal distributions (truncated at $\pm 2\sigma$): one with a mean of 20 feet per second (the hostile, or sn distribution); the other with a mean of 10 feet per second (the neutral, or n distribution). Both have a standard deviation of 2.5 feet per second. The new, actual altitude $a_i$ is then calculated from the descent rate and the time since the last probe $t_i$.

$$a_i = a_{i-1} - D_i t_i \qquad\qquad \text{(IV-4)}$$

The altitude displayed to the subject, $A_i$, is the actual altitude, plus an amount of noise $N_i$ sampled from a normal distribution (similarly truncated) with a mean of 0, and a standard deviation of 50 feet. An initial altitude at time ten seconds ($t_0$) is shown in addition to the altitude at the first probe ($t_0$).

The best strategy is to remember the initial altitude, and calculate a descent rate based on the current altitude, the initial altitude, and the time between them. Thus the estimate of descent rate for the $N^{th}$ probe, $\hat{d}_N$, is calculated as:

$$\hat{d}_N = \frac{A_N - A_0}{t_N} \tag{IV-5}$$

The variance of the initial displayed altitude, $V(A_0)$, is, of course, $50^2$ square feet. The variance of the $N^{th}$ displayed altitude, $V(A_N)$, is the sum of the variances of the altitude changes during the intervening probes, and the variance associated with the display of the altitude (i.e. $50^2$):

$$
\begin{aligned}
V(A_N) &= \sum_{i=1}^{N} t_i^2 V(D_i) + 50^2 \\
&= \sum_{i=1}^{N} (2.5 t_i)^2 + 50^2
\end{aligned}
\tag{IV-6}
$$

The variance of the estimated descent rate, $V(\hat{d}_N)$, is then the sum of the variances of the two altitudes from which it is calculated, divided by the square of the intervening time:

$$V(\hat{d}_N) = \frac{\sum_{i=1}^{N} (2.5 t_i)^2 + 2 \times 50^2}{(t_N - t_0)^2} \tag{IV-7}$$

In general, the interval of time between probes is constant, so a reasonable approximation gives:

71

$$\rightarrow V(\hat{d}_N) = \frac{\sum_{i=1}^{N} \frac{t_N^2}{N^2} 2.5^2 + 2 \times 50^2}{t_N^2} \, , \qquad t_i \approx \frac{t_N}{N}$$

$$= \frac{2.5^2}{N} + \frac{2 \times 50^2}{t_N^2} \, , \qquad t_N \approx N\bar{t} \qquad\qquad\qquad \text{(IV-8)}$$

$$\rightarrow V(\hat{d}_N) \approx \frac{2.5^2}{N} + \frac{2 \times 50^2}{N^2 \bar{t}^2} \qquad\qquad\qquad\qquad\qquad \text{(IV-9)}$$

Thus, since $d'$ is inversely proportional to the standard deviation of the estimate of descent rate, and the altitude rate variance term rapidly comes to dominate, this gives a graph of $d'$ against the $\sqrt{N}$ that is almost perfectly linear.

This strategy was taught to the subjects, which is not to say that it was used by all subjects. What subjects tended to do, rather than remember the first altitude and time, was to use the first displayed altitude and time, and the most recent altitude and time. There were five probe results displayed at a time. There were also some subjects who considered the concept too difficult, and so used the sub-optimal strategy that untrained subjects tended to use. In this strategy, the descent rate is calculated for each probe based on the altitude change since the last probe. This has the same variance as $\hat{d}_1$, and can even, with the high variance of altitude display, give an apparent altitude gain when two probes are taken close together. The altitude rates thus calculated were

then averaged[26] to arrive at a final conclusion. Optimal performance with this method, called the "sub-optimal observer," gives a smaller increase in $d'$ with probes:

$$V(\hat{d}_N) \approx \frac{(2.5\bar{t})^2 + 2\times 50^2}{\bar{t}^2 \times N}$$

$$= \frac{2.5^2 + 2\times\left(\dfrac{50}{\bar{t}}\right)^2}{N} \qquad\text{(IV-10)}$$

The difference is that the second term in equation (IV-10) is allowed to accumulate, where in equation (IV-9) it decayed rapidly with increasing trials because the variance of intervening altitude reports was not included in the final calculation.

### c.  Results for TIC

The performance of individual teams is shown in Figure 20a, with the expected performance of an ideal observer and a "sub-optimal observer." The ideal and sub-optimal lines are based on average times of probes for the number of probes, reduced by ten seconds to account for the fact that the initial probe reports $A_0$ for $t_0=10$ in addition to $A_1$. The results appear to cluster closely to the sub-optimal line, and start in general well below optimal. As with the IDS, the teams are combined by averaging individual $d''$'s, rather than using the collapsed $d'$. The average performance is shown in Figure 20b; again, a low start is followed by performance that tracks well with sub-optimal.

---

[26]In so far as any numerical results for the IDS and TIC were averaged. It is unlikely that symbolic arithmetic averaging took place at all: see earlier discussion on processes.
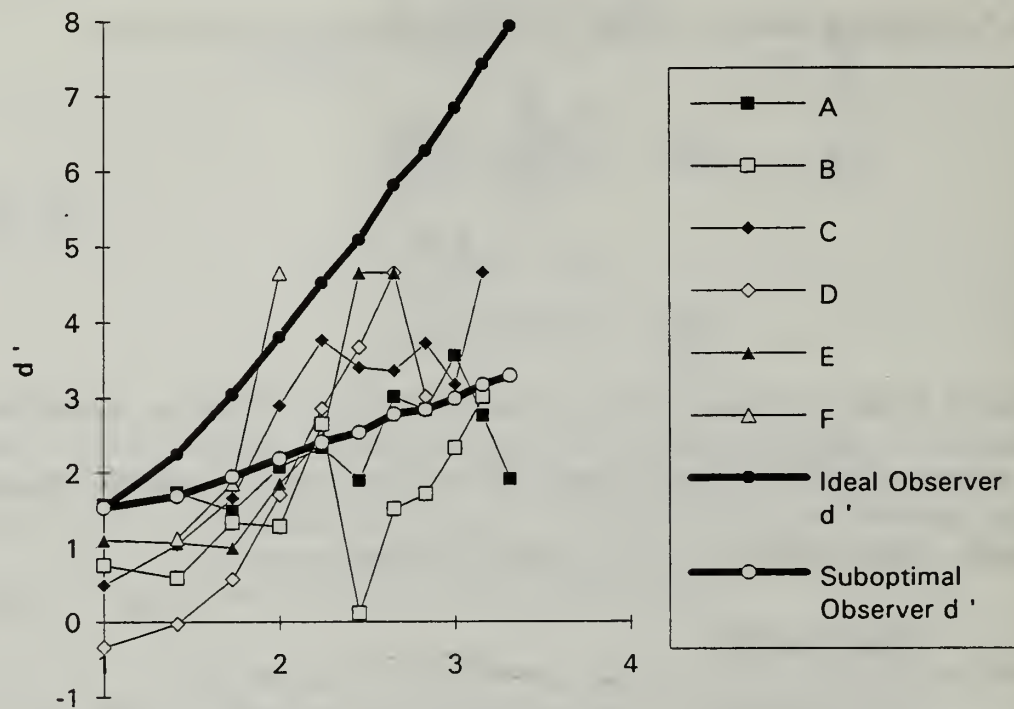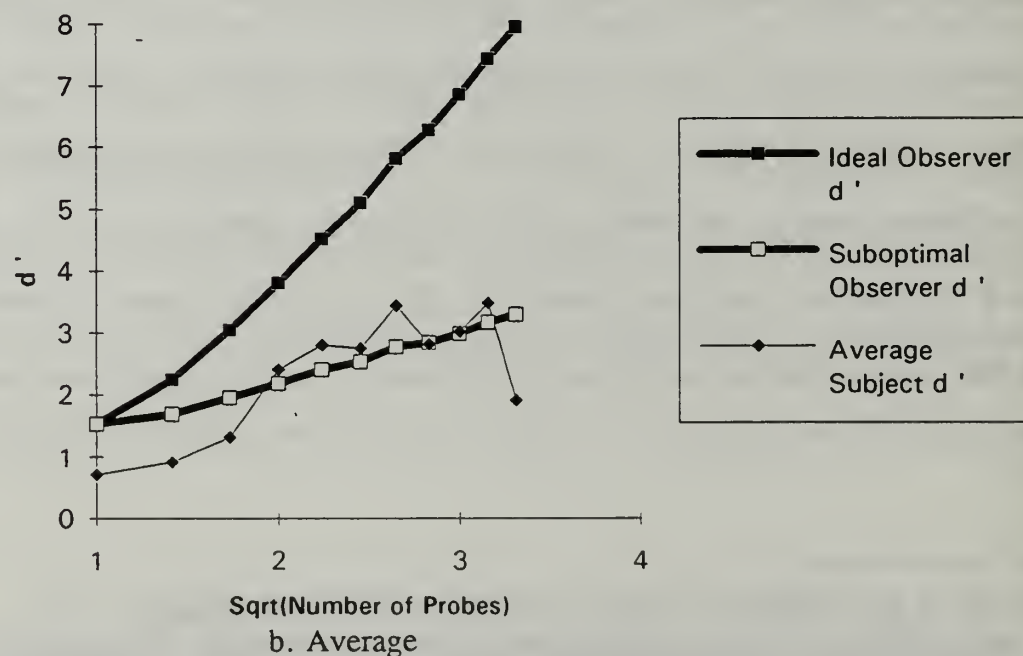
Figure 20: Team and Average Performance of TIC.

The reduced $d'$ on the first probe cannot indicate internal noise caused by arithmetic. To have sufficient noise (i.e., a sufficiently high standard deviation of a zero mean Gaussian noise distribution contributing extra variance to the measurement) to cause so severe a degradation, the internal noise variance would dominate throughout the experiment, and no improvement in $d'$ with $N$ would be seen. As with the IDS, some mechanism causing a reduced perceived difference between the means of the n and sn distributions would cause an overall decrement in $d'$. The result of halving the difference is shown in Figure 21. There is no simple, cognitive explanation for such a model. It is entirely possible, since in general the first few probes are unimportant except for gathering preliminary data, that players had a higher error rate because they simply tended to guess during the first two or three probes. It is also possible that the effect of using a $d'$ that is collapsed with respect to the independent variables, rather than averaged, causes a reduced result. This would be the case if the independent variables, in particular risk, caused markedly different operating points on the ROC for the same subject. Given the proximity of the first probe to ideal performance for the IDS, it is debatable whether collapsed $d'$ explains reduced performance for the IDS or TIC.

As noted in describing the ideal observer, most subjects did not remember the initial altitude, but used the maximum spread of readings available to them on the screen, which was a maximum of five. This would cap performance as shown in Figure 22a. The observed performance also matches the capped ideal observer with, again, a uniform decrement in $d'$ throughout the trial.
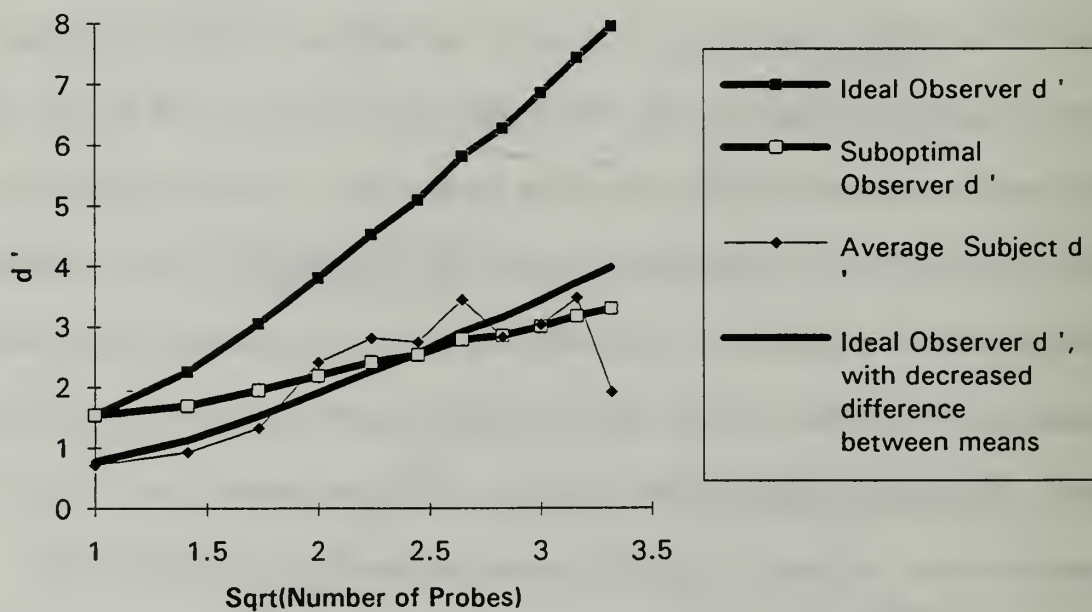
Figure 21: Reduced Difference Between Means Model for TIC.

Finally, it should be noted that where two strategies are available, different subjects will likely be using different strategies. Based on the differences between observed performance and optimal application of the two strategies, it appears possible that teams A and B used the sub-optimal strategy, while teams C through F used the ideal strategy. Separating these two groups gives the result shown in Figure 22b. Teams A and B track the sub-optimal observer with a small, uniform decrement; teams C through F track the ideal observer, probably with capped performance, with a larger decrement.

Lastly, the performance of $d'$ is compared again to average reported confidence. Figure 23a shows the two measures averaged for teams A and B, while Figure 23b shows them for teams C through F. The correlation is better in the latter case ($r=0.864$, significant at $p<0.01$) than the former ($r=0.642$, significant at $p<0.05$).

### 3. Electronic Warfare Supervisor (EWS)

#### a. *Proportion of Correct Assessments*

Figure 9 showed the EWS having an early peak, followed by a rapid decline to levels below the initial proportion of correct assessments. There appear in this graph to be severe limitations on the number of probes that can be successfully integrated. However, when the other stress levels are examined, as shown in Figure 24, it is apparent that significant numbers of probes can be successfully integrated, as demonstrated by medium stress trials. Evidently, there is a wrinkle present in the low stress data that suggests extreme caution in analyzing the data for the EWS.
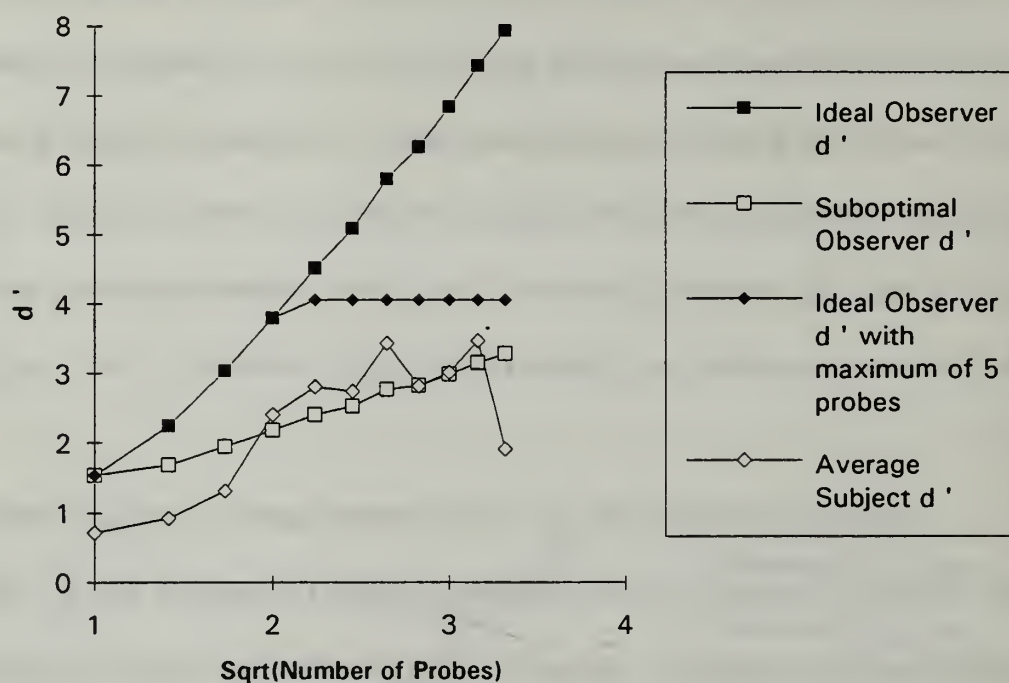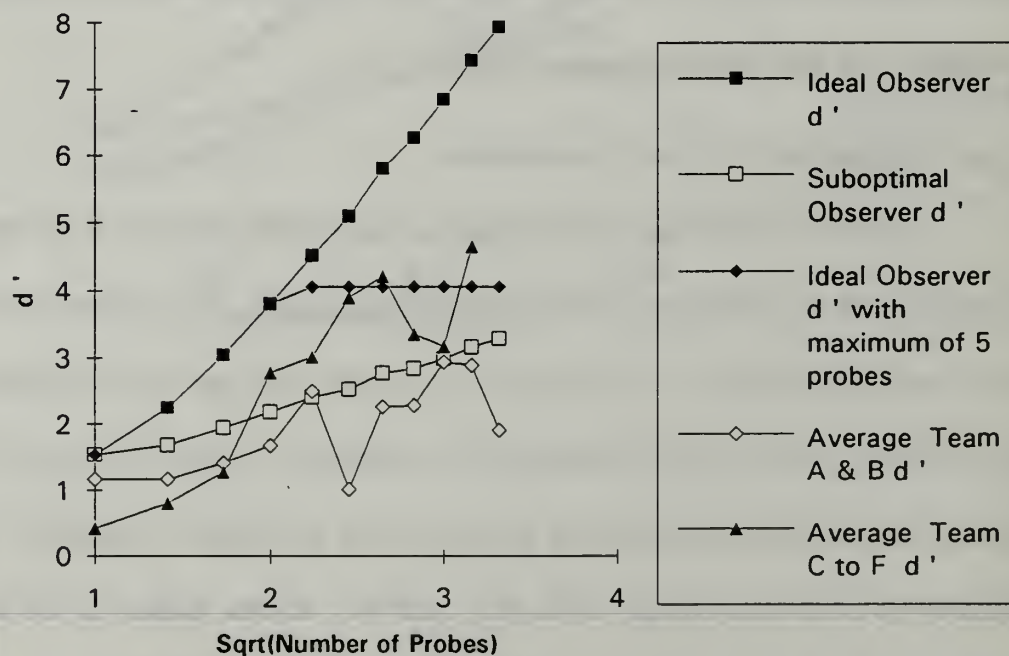
Figure 22: TIC Performance with Maximum of Five Probes for the Ideal Observer.
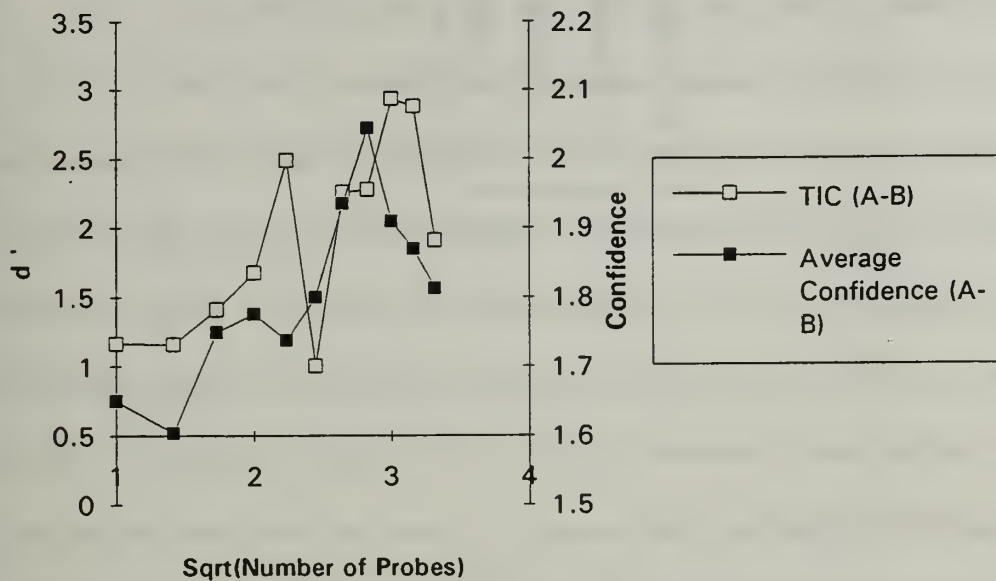
a. Teams A-B
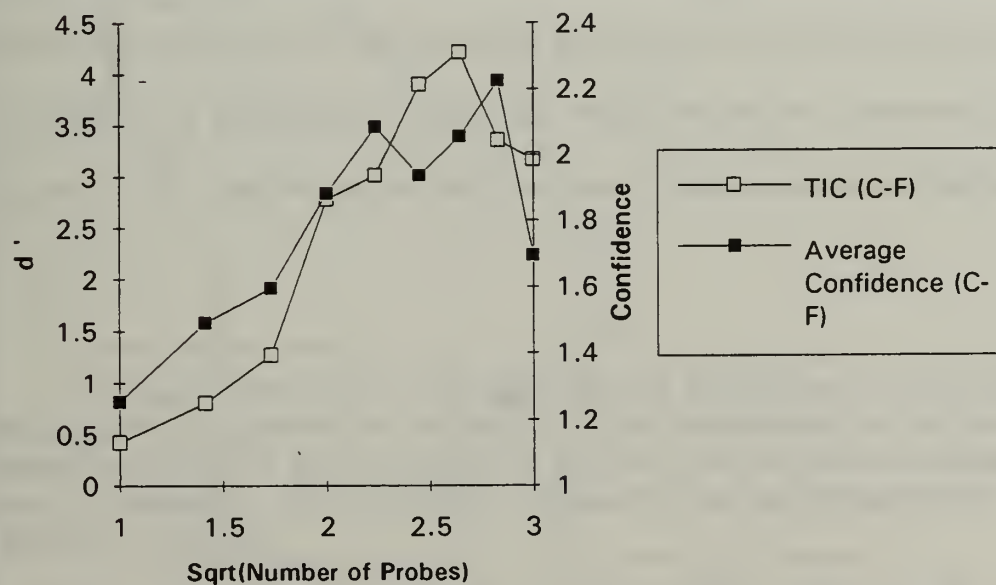


b. Teams C-F

Figure 23: Confidence Compared to $d'$ for TIC.
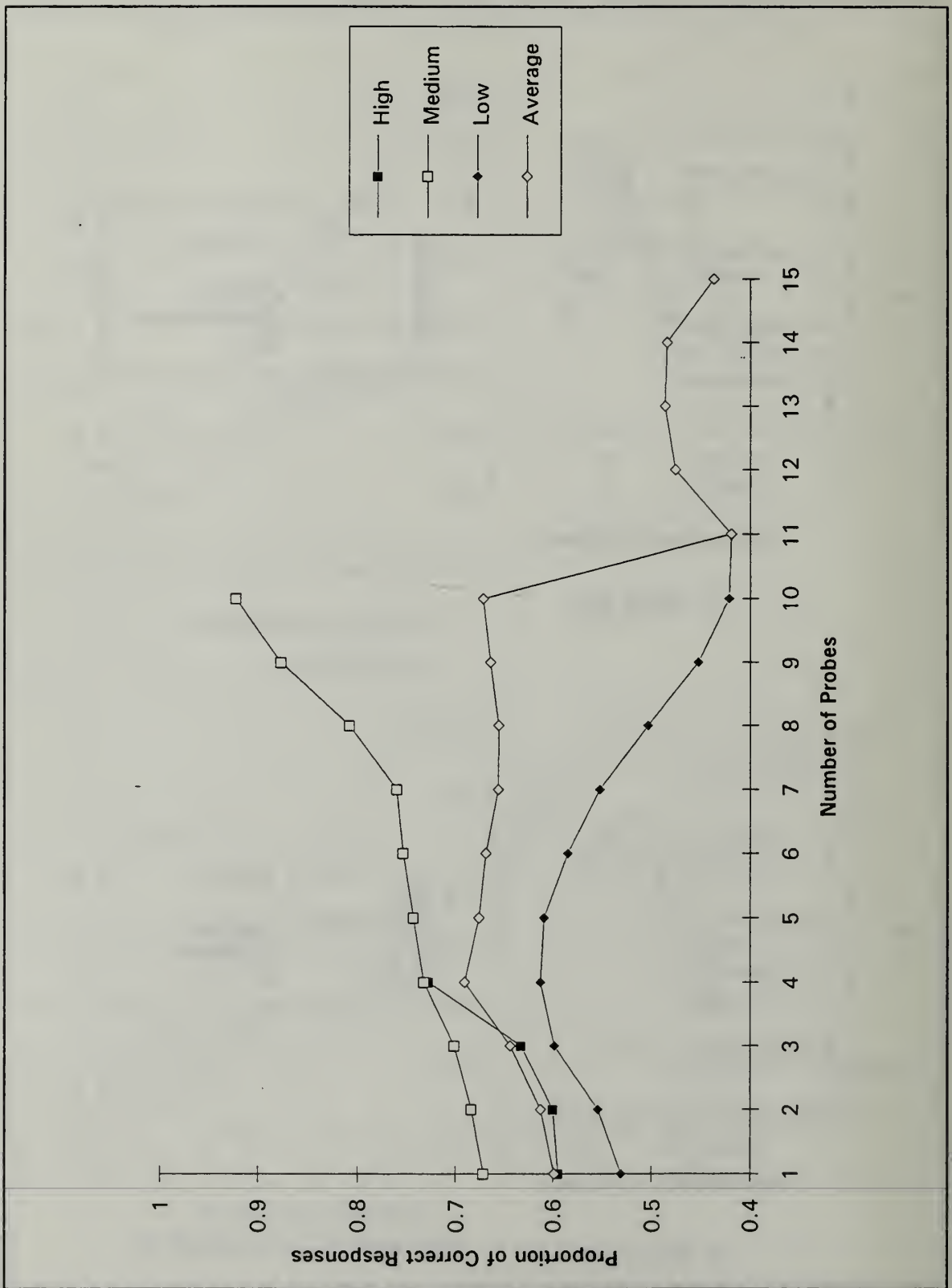
Figure 24: Proportion of Correct Assessments for EWS, by Level of Stress.

### b. Ideal Observer Performance

The EWS is at once the most difficult, and easiest, position for subjects to play. It is very easy, because the appropriate decision variable is the number of matches between the received pattern and the hostile pattern, an integer between 0 and 7, which merely needs to be accumulated across probes during the trial[27]. The cognitive demands of adding up a column of numbers between 0 and 7 are not great. However, the decision criterion is less clear. The IDS could readily use the value 50, and the TIC the value 30, as cutoff values[28], but for an EWS who accumulates matches, the cutoff value changes from probe to probe, in a manner that is not easy to calculate. Even on the first probe, since the mean depends on the probability of correct receipt of each bit, and this probability was not known to the subjects for the hostile distribution, an optimal criterion value could not be chosen. A very clever EWS with knowledge of the distributions would have written down the appropriate cutoff values before starting the trial session, for reference; but none of the subjects had access to the true distributions (at least of sn) or had the acumen and deviousness to devise such a stratagem.

---

[27]One subject (team F) was sometimes observed to take three or four probes without intervening assessment, particularly on low stress trials; then he would methodically calculate the total number of matches in all the probes displayed, and make an assessment. The other subjects appeared to use an approach more "Bayesian" in nature.

[28]These values would be modified in high risk trials by an ideal observer. However, there is evidence that human subjects are very poor at conducting an optimal likelihood ratio test where prior probabilities and/or payoffs are not symmetrical, e.g., Kubovy & Healy (1980).

The n and sn distributions in the case of the EWS are neither normal, nor equivariant. The n distribution, representative of a neutral, is a binomial, with p=0.5, and seven Bernoulli trials per probe. Thus, for N probes, the mean of the n distribution is $3.5 \times N$, and the variance is $1.75 \times N$. The sn consists of the same number of Bernoulli trials, but with p=0.7. Thus its mean is $4.9 \times N$, and its variance $1.47 \times N$. The way to assess $d'$ for distributions that are not equivariant is to measure the distance between the means in units of the standard deviation of the n distribution. For observed data, $d'$ is then given by:

$$d' = \frac{\sigma_n}{\sigma_{sn}} z[P(\text{Hit}|\text{sn})] - z[P(\text{False Alarm}|\text{n})] \qquad \text{(IV-11)}$$

(Macmillan & Kaplan, 1985). This assumes that the deviations of the internal representation of the n and sn distributions are at least in the same proportion as the actual distributions; an assumption that would be difficult to justify.

### c. Results for EWS

The variation with $\sqrt{N}$ of the ideal observer $d'$ is shown in Figure 25. Also shown are the observed values for the teams individually, and the average of the observed values. Note that the dip in performance after eight probes is due entirely to the spate of appalling performance by team D's EWS. These probes are below the chance level, and can only represent confusion. This single person's confusion accounts for the dip in $d'$ and, presumably, a large part of the fall in proportion of correct assessments for between eight and twelve probes (see Figure 24).

82
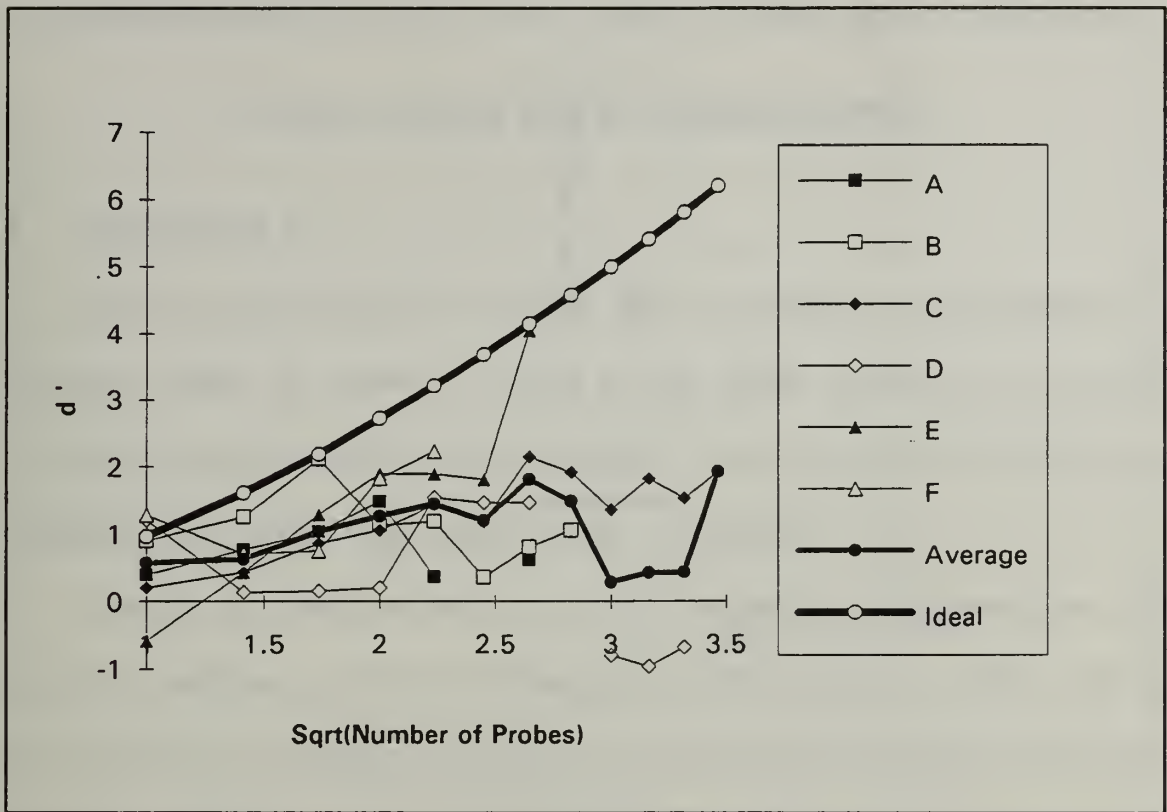
Figure 25: EWS Performance as a Function of Number of Probes; Ideal Observer, by Team, and Average

The rest of Figure 25 indicates performance that is at least uniformly below optimal, and generally increasingly so with more probes. This latter behavior has been seen to be indicative of an extra source of (internal) noise, probably from the demands of arithmetic.

The relationship between $d'$ and average reported confidence for the EWS is shown in Figure 26. The match is reasonably good (r=0.670, p<0.02), the dip in confidence probably reflecting the confusion of the team D subject.
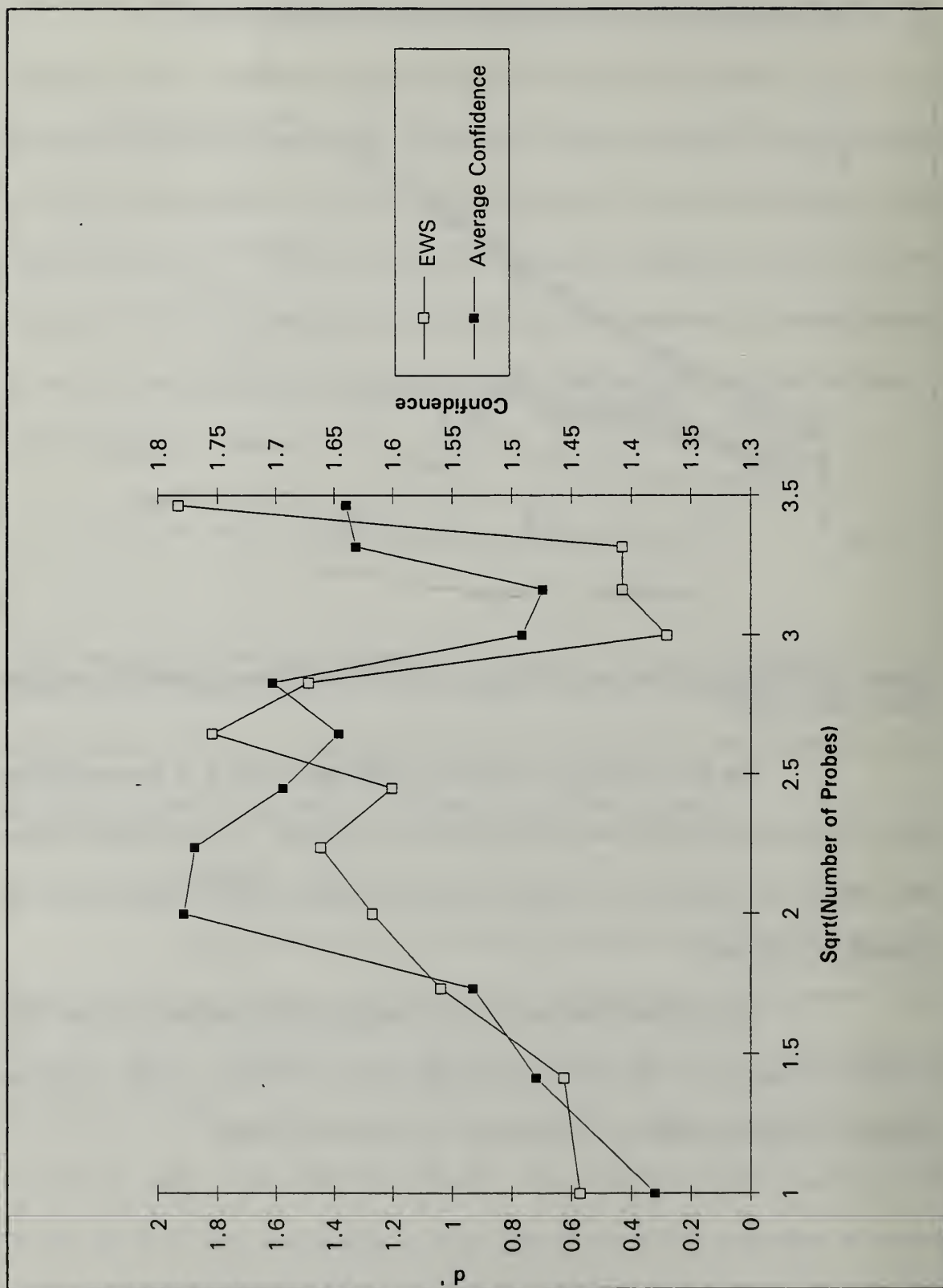
Figure 26: EWS Performance Compared to Average Confidence.

# V. CONCLUSIONS AND RECOMMENDATIONS

## A. DEPRESSED $d'$

Two observations were made repeatedly: subject performance did not improve with increasing number of probes as rapidly as the optimal observer's, and subject performance was consistently lower than optimal. In some cases, the data best fit one observation or the other, rather than both--e.g., the IDS data.

There are two possible methods for the subjects to integrate their observations, and thus two possible explanations for a reduced rate of improvement in $d'$. First, subjects may average (or integrate--the strategies are equivalent) the observations, and compare the average to a criterion value (e.g., 50 for the IDS, 15 for the TIC, 4 for the EWS). This calculation is one potential source of errors. Alternatively, the subjects may calculate a likelihood ratio, and accumulate the ratios (multiplicatively) as more probes are obtained.[29] This can also introduce errors, since the subjects did not know the variances of the distributions concerned.

---

[29]This fits informal observation of the subjects: a typical comment by a subordinate might be: "This reading is slightly hostile, but the last two readings were very neutral, although the one before that was extremely hostile, so I will guess hostile, with low confidence." Words like "slightly," "very," and "extremely" express likelihood ratios, which the subject is clearly trying to average in some, non-numeric, way.

Randomness in the calculated likelihood ratio, or the criterion used in a LRT,[30] does not, in itself, change the value of $d'$; rather it makes the operating point on the ROC random, which can reduce below optimal the average score or percentage of correct assessments. However, the calculated $d'$ used in this study is not the average of the $d'$ for each trial for the given observer. Because the hit and false-alarm probabilities were calculated for many trials, even if for the same observer, effectively a collapsed $d'$ has been obtained. If the one observer then uses very different criteria from trial to trial, the calculated $d'$ will be reduced (Macmillan and Kaplan, 1985). It is thus possible that the effect of lowered $d'$ being seen in this study and elsewhere (see, for example, the ROCs obtained in Pete, Pattipati, & Kleinman, 1993:1) are not representative of cognitive limitations, but only of collapsing data across random criteria.

The solution, as is so often the case, is to increase the amount of data collected. With more trials under the same conditions, the experimenter can be more confident that the variance in the comparison criterion used is smaller than when data must be collapsed across different conditions. Were this done, a probe-by-probe comparison of $d'$ for the different conditions of the independent variables could be accomplished.

## B. IMPORTANCE OF STRATEGIES OF SUBJECTS

The CHIPS experiment is part of a normative-descriptive effort, in which the performance of human subjects is compared to the optimal performance predicted by the

---

[30]These are equivalent: whether the randomness is on one side of the comparison or the other the effect is the same.

normative model. The comparison allows the calculation of values of parameters that modify the normative model to account for human cognitive limitations (Pete, Pattipati, & Kleinman, 1993:1).

To be able to determine how to modify the normative model, it is important to understand the cognitive strategy, or *process*, being used by the subjects (Payne, 1980).

> Payne is concerned with what information the subject uses and with how that information is used. He guesses that a situation is often processed differently by the observer than is presumed by the experimenter. He conjectures that, by a process-tracing procedure, one might learn what information is being used. (Lockhead, 1980)

While parameters may be found that can be tuned to provide very accurate fits with observed data, this does not necessarily provide accurate indications on how to improve performance. It is important that the processes being used in making decisions are known, and then compared to optimal performance with this same process. This provides two methods for improving performance: either a different process can be trained, if optimal performance with the process currently being used falls significantly below normative performance; or a method can be found to improve the use of the current process, such as providing computer assistance to overcome cognitive limitations of the human decision-maker.

## C. POTENTIAL IMPROVEMENTS IN EXPERIMENT DESIGN

### 1. Determination of Process

There are two ways to determine the process being used by subjects. One is to require subjects to describe what they are thinking throughout the experiment.

Either this description must be monitored by an observer with sufficient skill and knowledge to assess the process being used (perhaps from a menu of potential processes, based on pre-established criteria), or it must be recorded, and later analyzed. Alternatively, the process to be used can be imposed on the subject. For example, comparison to a specific value may be required, and the subject told to record the calculated average of observed values. To compare the imposed process to the processes naturally selected by subjects would require use of a control and an experimental group: the control group would spend two, two-hour sessions with the simulator, both with no process imposed; the experimental group would spend one two-hour session with the simulator without an imposed process, then be trained on the method to be tested, and given another two-hour session. If the imposed process is superior to the subject's own, then the experimental group should show significantly greater improvement in the second session than the control group.

### 2.    Recording of Data

There were strong indications of recency effects during the experiment--one TIC, during training, would, for example, alternate with each probe between high confidence neutral and high confidence hostile, based entirely on the most recent probe. However, the presence of recency effects could not adequately be tested because the indications given to the subject were not recorded in the Log File.

### 3. Generation of Stress

The fact that fewer probes were made during higher stress trials confounds the investigation of the effect of stress. There are other ways to increase stress without necessarily changing the probe rate or total number of probes. Increasing the number of targets of interest would increase the stress without changing the length of a trial, but would most likely reduce the number of probes per target. A distractor task could be introduced, and made progressively more demanding, although again this could reduce probe rate. One method that would not reduce the probe rate is to limit the duration that probe measurements remain visible, and shorten the duration for higher stress levels.

There are also ways to ensure that probe rate is not greatly affected, even if one of the first two methods is used. Probing could become an automatic function, so that the subject need only make an assessment each time he is presented with a probe. Of course, this would remove probe rate as a dependent variable.

### 4. Change in Computer Screen Layout

The graphics display of aircraft, which is a vital element in DDD-II, is completely irrelevant in CHIPS, except insofar as it distinguished between subjects on the basis of manual dexterity with a mouse. The window displaying the results of probes could remain open continuously, if probe information were allowed to appear in it without having to close and re-open the window.

Similarly, the TAO fusion window could remain open if subordinate assessments were able to appear in the open window. This would have the distinct advantage that the information being used by the TAO to make his assessment would be

the same as the information that the computer recorded as being available. This could be further enhanced: just as it would be desirable for the Log File to record the measurements presented to the subordinates on which a decision was made, it would also be desirable for the assessments shown to the TAO to be recorded with his fused assessment based on them.

The problem of conflicting information available to the TAO between that recorded in the computer and that reported over the intercom could be avoided in one of two ways--by disabling one or the other input. If the intercom were removed, the problem would similarly be removed. TAO update could still be studied as an independent variable, by having the computer report the TAO's assessment to the subordinates each time one is made during update trials. Alternatively, the automatic presentation of subordinate assessments to the TAO could be removed, requiring him instead to record on his fusion screen the latest (verbal) report from each subordinate, as a part of making his assessment.

## D.   CONCLUSIONS

It was found that subordinates were less able to distinguish between hostile and neutral indications than the optimal observer, and that the discrepancy tended to increase with time. This could be attributed to: arithmetic capability limitations of subjects, poor judgment of distribution variance, or simply to the fact that the measure used is artificially lowered when probabilities are averaged for observations using significantly different criteria. Further work is needed to develop a method for determining the

cognitive process employed by decision-makers, so that they may be replaced or supplemented to improve performance. While the processes used by subjects in the abstraction of a command center present here are not necessarily relevant to any military application, the methods by which they are determined could be applied in assessing and improving actual command centers.

# APPENDIX A: STATISTICAL ANALYSES

## A. RESULTS BY TEAM

### 1. Comparison to Chance Performance

#### a. *Improvement of Performance During the Trial -- The McNemar Test*

Final assessments by each team were compared to initial assessments to determine whether the extra time spent probing and reporting on the target contributed to extra correct assessments. A simple comparison of proportions (or numbers) (e.g., by the large sample normal approximation for comparison of proportions) would be unrevealing in this case. However the data are naturally paired, leading to the McNemar test, as described in Pratt & Gibbons (1981, p. 108).

The McNemar test examines cases in which subjects are classed by a dichotomous test (e.g., right/wrong assessment) twice--once before and once after a treatment. The null hypothesis tested is that the treatment has no effect on the outcome of the test. Four counts are determined: (A) the number of subjects who were, for example, wrong on both tests, (B) the number who changed from wrong to right, (C) the number who changed from right to wrong, and (D) the number who were right on both tests. The first and last counts (A and D) are, for this test, irrelevant. Now, the null hypothesis can be equivalently stated as: of those subjects who were right on one test and wrong on the other, the probability that they changed from wrong to right vice right to wrong is 0.5. This is tested simply using the Binomial test (either one-tailed or two-

tailed, as appropriate for the theory being tested): if the smaller of B or C (as appropriate for the tail being tested, for a one-tailed test) is less than or equal to the critical value for a (lower- or two-tailed) Binomial test, where the number of trials is the sum of B and C, then the null hypothesis is rejected.

In the present case, a one-tailed test is to be performed, looking for significantly more trials which turned an initial guess that was wrong into one that was right than vice versa. Since teams A, B and F all declined in performance, and C showed no change, no further investigation of these teams is required. For D and E, the results are summarized in Table VI, showing significant effect in both cases.

TABLE VI: MCNEMAR TEST FOR IMPROVED PERFORMANCE

| Team D | | Initial Assessment | |
| --- | --- | --- | --- |
| | | Wrong | Right |
| Final Assessment | Wrong | 6 | 2 |
| | Right | 5 | 11 |
| Critical value (p=0.1): 2 | | | |

| Team E | | Initial Assessment | |
| --- | --- | --- | --- |
| | | Wrong | Right |
| Final Assessment | Wrong | 5 | 4 |
| | Right | 8 | 7 |
| Critical value (p=0.1): 4 | | | |

### b. *Effect of Team on Time and Probes at First Assessment*

The significance of the differences in times to first assessment by the six teams was tested with ANOVA, using Tukey's procedure to determine pairs of teams with significantly different times. The Minitab output is shown in Table VII.

TABLE VII: VARIANCE OF TIME TO FIRST ASSESSMENT WITH TEAM

```
ANALYSIS OF VARIANCE ON time
SOURCE      DF        SS        MS       F        p
team         5     7204.8    1441.0   34.84    0.000
ERROR      177     7321.4      41.4
TOTAL      182    14526.2
                                  INDIVIDUAL 95 PCT CI'S FOR MEAN
                                  BASED ON POOLED STDEV
  LEVEL      N      MEAN     STDEV  -------+---------+---------+---------
    1       32    39.563     7.670  (---*--)
    2       30    46.300     6.803             (--*--)
    3       31    40.194     5.443   (--*---)
    4       31    43.032     5.250        (--*---)
    5       32    42.094     7.222       (--*--)
    6       27    58.778     5.625                         (---*--)
                                  -------+---------+---------+---------
POOLED STDEV =      6.431          42.0      49.0      56.0


Tukey's pairwise comparisons

     Family error rate = 0.0500
Individual error rate = 0.00445

Critical value = 4.07

Intervals for (column level mean) - (row level mean)

              1         2         3         4         5

    2    -11.441
          -2.034

    3     -5.296     1.366
           4.033    10.847

    4     -8.134    -1.473    -7.540
           1.195     8.008     1.863

    5     -7.159    -0.498    -6.565    -3.726
           2.096     8.910     2.764     5.603

    6    -24.052   -17.388   -23.457   -20.618   -21.521
         -14.378    -7.568   -13.712   -10.873   -11.847
```

94

The same method was used to analyze the data for number of subordinate probes. Again, Minitab output is shown in Table VIII.

## TABLE VIII: VARIANCE OF NUMBER OF PROBES WITH TEAM

```
ANALYSIS OF VARIANCE ON NumProbe
SOURCE     DF        SS        MS        F        p
Team        5    193.97     38.79    17.19    0.000
ERROR     177    399.45      2.26
TOTAL     182    593.42
                                    INDIVIDUAL 95 PCT CI'S FOR MEAN
                                    BASED ON POOLED STDEV
  LEVEL     N      MEAN     STDEV   --------+---------+---------+--------
      1    32     7.844     1.780             (---*----)
      2    30     9.533     1.676                              (---*----)
      3    31     6.871     1.147   (---*----)
      4    31     8.645     1.170                      (---*---)
      5    32     6.844     1.725   (---*---)
      6    27     9.037     1.344                         (---*----)
                                    --------+---------+---------+--------
POOLED STDEV =     1.502               7.2       8.4       9.6

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.00445

Critical value = 4.07

Intervals for (column level mean) - (row level mean)

              1         2         3         4         5

    2    -2.788
         -0.591

    3    -0.117     1.555
          2.062     3.770

    4    -1.891    -0.219    -2.872
          0.288     1.995    -0.676

    5    -0.081     1.591    -1.062     0.712
          2.081     3.788     1.117     2.891

    6    -2.323    -0.651    -3.304    -1.530    -3.323
         -0.064     1.643    -1.028     0.746    -1.064
```

c. *Relation Between Time at First Assessment and Number of Correct Initial Assessments*

Three methods were used to attempt to demonstrate a relation between the average time at which the TAO makes his first assessment, and the number of correct initial assessments by the TAO. The Pearson correlation coefficient between the time and number was calculated, giving $r=0.43$. To be significant at $p=0.1$, with 4 degrees of freedom (sample size - 1), r would have to be at least 0.729 (Snedecor and Cochran, 1980, p. 477). There is no reason to believe that the distribution of either variable is normal, however; so the use of the Pearson correlation coefficient is suspect. Indeed, the distribution of the time at first assessment is positively (in the technical sense) skewed (see Figure 27 for the distribution, and ? for the normal probability plot). Consequently, the correlation between the ranks of the teams based on time at, and success of, first probe was calculated, giving the Spearman rank correlation coefficient. In this case, $r_s=0.40$. To be significant at $p=0.1$, with sample size 6, requires $r_s$ be at least 0.771.
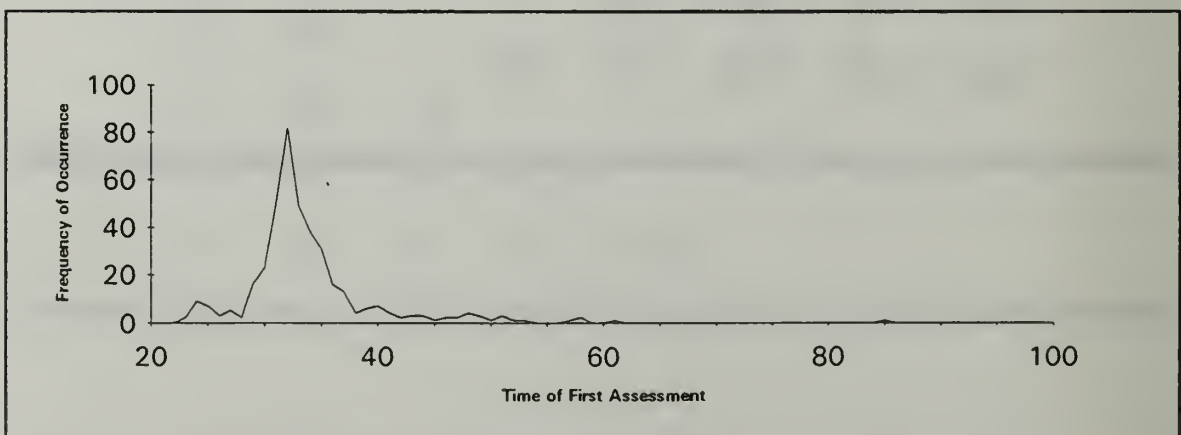


Figure 27: Distribution of Time to First Assessment.

```
MTB > Plot 'Time1st' 'N-score';
SUBC>    Symbol 'x'.


 Time1st -
         -                                                               x
         -
    75+
         -
         -
         -                                                         2  x
         -                                                        xx
    50+                                                     7x3x
         -                                                362
         -                                              +76
         -                                    +    +  ++   +
         -                      2+   +    +       +
    25+          2      9    735
         -
         -
         -

- - - - - - - +- - - - - - - - +- - - - - - - - +- - - - - - - - +- - - - - - - - +- - - - - - - N-score
             -2.4         -1.2         0.0          1.2          2.4
```
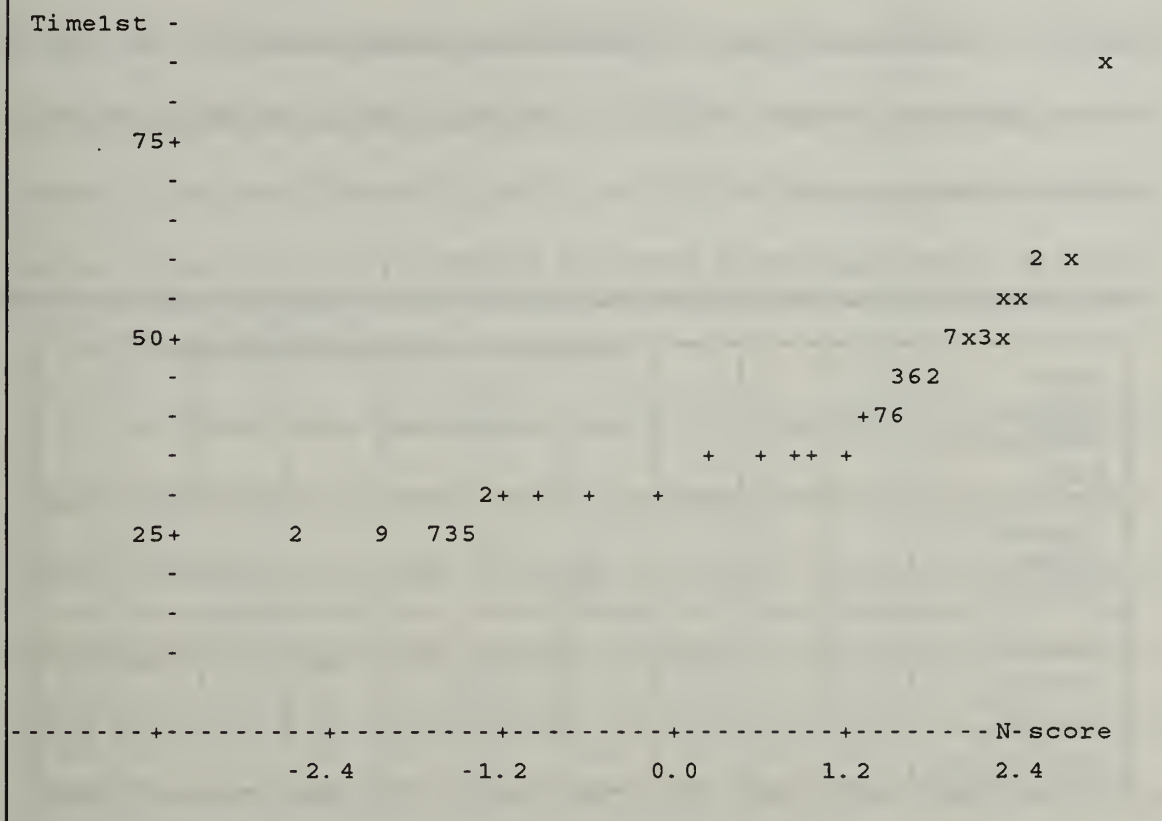
Figure 28: Normal Probability Plot of Time to First Assessment.

Lastly, a randomization method was used. Given a set of six teams ranked one to six on one measure, there are only 720 different ways in which the teams may be ranked on a second measure[31]. The degree of similarity in the two rankings is expressed by the sum of the absolute differences in rank between the two measures,

---

[31]This presentation neglects the possibility of ties. There are several methods available for handling ties. In the present treatment, each of the possible assignments of the affected ranks to the teams was tried, and the largest difference calculated by this method used in the analysis. This is a conservative method. For example, if the ranks were 1, 2=, 2=, 4, 5=, 5=: the rankings tried were 1, 2, 3, 4, 5, 6; 1, 2, 3, 4, 6, 5; 1, 3, 2, 4, 5, 6; and 1, 3, 2, 4, 6, 5.

across the six teams. The 720 possible random assignments are easily enumerated (a BASIC language program was used) and the sum of absolute differences calculated (see Table IX). From these values, a cumulative distribution function for the Sum of Absolute Differences of Rank statistic is readily calculated for the null hypothesis of random distribution of ranks.

TABLE IX: RANDOMIZATION TEST FOR RANKS 1 TO 6

| Sum of Absolute Differences of Rank | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|
| Frequency of Random Occurrence | 1 | 5 | 18 | 46 | 93 | 137 | 48 | 100 | 100 | 36 |
| Cumulative Frequency | 1 | 6 | 24 | 70 | 163 | 300 | 448 | 584 | 684 | 720 |
| Cumulative Probability (%) | 0.1 | 0.8 | 3.3 | 9.7 | 23 | 42 | 62 | 81 | 95 | 100 |

The critical values are then read (approximately) from the table. For $p = 0.1$, any value of 6 or less is significant. At $p = 0.05$ (actually 0.033), any value of 4 or less is significant. At $p = 0.01$ (actually 0.008), any value of 2 or less is significant.

For the relation between time of first assessment and number of correct initial assessments, the sum of absolute differences of rank calculated is 8 or 10, depending on the assignment of ranks to the two ties. Thus this statistic also fails to demonstrate a significant relationship.

### d. Relation Between Number of Subordinate Probes at First Assessment and Number of Correct Initial Assessments

The Pearson Correlation Coefficient r=0.88, which is significant at p=0.05. The Spearman Rank Correlation Coefficient $r_s$=0.83, significant at p=0.1. The sum of absolute differences of rank is 2, 4 or 6 depending on the random assignment for ties, giving p<0.1.

### 2. Effect of Independent Variables

The McNemar test was again used to assess the effect of each independent variable individually (see p. 92 above for a description of this test). The pairing of the data is less obvious in this case. Consider, for example, the case of TAO feedback: each team saw 24 balanced trials (and eight distractor trials, omitted from this analysis), in half of which the TAO provided updates, and half of which he did not. In each group of twelve, every combination of risk (two levels), stress (three levels), and contact classification (two levels: neutral or hostile) was seen exactly once. Therefore the trials can be paired based on these three factors (plus team), and the effect of TAO updates examined within the pairs. The results are presented in Table X.

## B. RESULTS FOR INDIVIDUAL SUBORDINATE ROLES

### 1. Effect of Stress Level on Variation of Proportion of Correct Assessments with Number of Probes

The proportions of correct assessments were transformed with the angle transformation for proportions (arc sin($\sqrt{p}$)), and analyzed with ANOVA and Student's

# TABLE X: MCNEMAR TESTS FOR EFFECT OF INDEPENDENT VARIABLES

| **TAO Update:** | | No update | | Level of Significance: |
|---|---|---|---|---|
| | Number of Assessments | Wrong | Right | p=0.15 |
| Update | Wrong | 15 | 19 | |
| | Right | 14 | 24 | |

| **Risk:** | | Low | | Level of Significance: |
|---|---|---|---|---|
| | Number of Assessments | Wrong | Right | p=0.1 |
| High | Wrong | 14 | 21 | |
| | Right | 14 | 23 | |

| **Stress:** | | Low | | Level of Significance: |
|---|---|---|---|---|
| | Number of Assessments | Wrong | Right | p=0.1 |
| Medium | Wrong | 8 | 13 | |
| | Right | 7 | 20 | |

| | | Medium | | Level of Significance: |
|---|---|---|---|---|
| | Number of Assessments | Wrong | Right | p=0.1 |
| High | Wrong | 11 | 16 | |
| | Right | 10 | 11 | |

| | | Low | | Level of Significance: |
|---|---|---|---|---|
| | Number of Assessments | Wrong | Right | p=0.1 |
| High | Wrong | 11 | 16 | |
| | Right | 4 | 17 | |

t-test. Because the number of values at each level of stress are different (there were a maximum of four probes for which useful proportions could be calculated at high stress, ten at medium stress, and 15 at low stress), analysis of variance had to be performed separately for the first four probes, using all stress levels, and for the first ten probes, using only medium and low stress. The t-test was performed between pairs of stress levels.

### a. Results for IDS

The t-tests and results of ANOVA for the IDS are shown in Tables XI, XII and XIII. There are significant differences between the proportions, particularly between low and medium stress trials. Meanwhile, the number of probes does not appear to have a significant effect on the proportion of correct assessments, at least during the first ten probes.

TABLE XI: SIGNIFICANCE LEVELS OF T-TESTS BETWEEN PAIRS OF STRESS LEVELS FOR IDS

| p-values | Low | Medium |
|----------|-------|--------|
| High | 0.347 | 0.078 |
| Medium | <0.01 | |

# TABLE XII: ANOVA FOR IDS TRANSFORMED PROPORTIONS OF CORRECT ASSESSMENTS, ALL LEVELS OF STRESS, FIRST FOUR PROBES

Anova: Two-Factor Without Replication

|  | Summary | Count | Sum | Average | Variance |
|---|---|---|---|---|---|
| Stress: | High | 4 | 247.829 | 61.9573 | 52.9902 |
|  | Medium | 4 | 215.766 | 53.9416 | 0.31434 |
|  | Low | 4 | 218.103 | 54.5258 | 13.588 |
| Number of probes: | 1 | 3 | 162.959 | 54.3198 | 12.6665 |
|  | 2 | 3 | 167.696 | 55.8987 | 21.4449 |
|  | 3 | 3 | 166.426 | 55.4754 | 1.1501 |
|  | 4 | 3 | 184.617 | 61.5391 | 98.1918 |

ANOVA

| Source of Variation: | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| Rows | 159.759 | 2 | 79.8797 | 4.47308 | 0.06469 | 5.14325 |
| Columns | 93.5305 | 3 | 31.1768 | 1.74583 | 0.2569 | 4.75706 |
| Error | 107.147 | 6 | 17.8579 |  |  |  |
| Total | 360.437 | 11 |  |  |  |  |

## TABLE XIII: ANOVA FOR IDS TRANSFORMED PROPORTIONS OF CORRECT ASSESSMENTS, LOW AND MEDIUM STRESS, FIRST TEN PROBES

Anova: Two-Factor Without Replication

| | Summary | Count | Sum | Average | Variance |
|---|---|---|---|---|---|
| Stress Level: | Medium | 10 | 524.079 | 52.4079 | 3.09984 |
| | Low | 10 | 581.312 | 58.1312 | 14.4908 |
| Number of Probes: | 1 | 2 | 105.038 | 52.5191 | 5.87767 |
| | 2 | 2 | 106.55 | 53.2748 | 1.58174 |
| | 3 | 2 | 110.116 | 55.0581 | 1.25555 |
| | 4 | 2 | 112.166 | 56.0828 | 17.7583 |
| | 5 | 2 | 113.409 | 56.7047 | 36.4692 |
| | 6 | 2 | 112.597 | 56.2985 | 52.4689 |
| | 7 | 2 | 110.835 | 55.4177 | 64.6846 |
| | 8 | 2 | 110.416 | 55.2081 | 55.5344 |
| | 9 | 2 | 110.97 | 55.4852 | 34.2719 |
| | 10 | 2 | 113.293 | 56.6466 | 17.5175 |

ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| Rows | 163.778 | 1 | 163.778 | 11.9215 | 0.00724 | 5.11736 |
| Columns | 34.6739 | 9 | 3.85266 | 0.28044 | 0.96401 | 3.1789 |
| Error | 123.642 | 9 | 13.738 | | | |
| Total | 322.094 | 19 | | | | |

## 2. Fitting Regression Lines to Observed Data

### a. IDS

Linear regression was used to fit the lines shown in Figure 15 (and following figures). The results of fitting the ideal observer points are shown in Table XIV. The line can be seen to have a slope of about 0.86.

TABLE XIV: REGRESSION ANALYSIS FOR IDEAL IDS

| Regression Statistics | Ideal Observer |
|---|---|
| Multiple R | 0.99921 |
| R Square | 0.99843 |
| Adjusted R Square | 0.99823 |
| Standard Error | 0.02568 |
| Observations | 10 |

Analysis of Variance

| | df | Sum of Squares | Mean Square | F | Significance F |
|---|---|---|---|---|---|
| Regression | 1 | 3.35627 | 3.35627 | 5087.8 | 1.7e-12 |
| Residual | 8 | 0.00528 | 0.00066 | | |
| Total | 9 | 3.36154 | | | |

| | Coefficients | Standard Error | t Statistic | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 0.16734 | 0.02834 | 5.90482 | 0.00023 | 0.10199 | 0.23269 |
| x1 | 0.86193 | 0.01208 | 71.3288 | 1.1e-13 | 0.83406 | 0.8898 |

104

The results for fitting the observed data, without constraining the intercept, are shown in Table XV. The slope of the ideal observer is seen to be within the 95% confidence limits for the slope of the regression line, so the hypothesis that the slopes are the same is not rejected at the 5% significance level.

TABLE XV: UNCONSTRAINED REGRESSION OF OBSERVED IDS DATA

| Regression Statistics | Observed |
|---|---|
| Multiple R | 0.90939 |
| R Square | 0.82698 |
| Adjusted R Square | 0.80536 |
| Standard Error | 0.2457 |
| Observations | 10 |

Analysis of Variance

| | df | Sum of Squares | Mean Square | F | Significance F |
|---|---|---|---|---|---|
| Regression | 1 | 2.30834 | 2.30834 | 38.2387 | 0.00026 |
| Residual | 8 | 0.48293 | 0.06037 | | |
| Total | 9 | 2.79127 | | | |

| | Coefficients | Standard Error | t Statistic | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 0.13333 | 0.2711 | 0.4918 | 0.63463 | -0.4918 | 0.75847 |
| x1 | 0.71482 | 0.1156 | 6.18375 | 0.00016 | 0.44825 | 0.98138 |

To constrain the regression line to pass through the initial ideal $d'$ the axes were shifted, and a regression line fitted through the origin. The result is shown in Table XVI. The intercept is meaningless, since the axes have not been shifted back. However, the slope is meaningful, and the fact that the value of the ideal observer slope does not fall within the 99% confidence interval indicates that the two slopes are different, significant at the 0.01 level.

## TABLE XVI: REGRESSION FOR IDS DATA, CONSTRAINED TO PASS THROUGH INITIAL IDEAL POINT

| | Regression Statistics | Through $d'_1$ |
|---|---|---|
| Multiple R | 0.90055 | |
| R Square | 0.81099 | |
| Adjusted R Square | 0.69988 | |
| Standard Error | 0.24211 | |
| Observations | 10 | |

Analysis of Variance

| | df | Sum of Squares | Mean Square | F | Significance F |
|---|---|---|---|---|---|
| Regression | 1 | 2.26371 | 2.26371 | 38.6174 | 0.00026 |
| Residual | 9 | 0.52757 | 0.05862 | | |
| Total | 10 | 2.79127 | | | |

| | Coefficients | Standard Error | t Statistic | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 0 | 0 | 0 | 0 | 0 | 0 |
| x1 | 0.62732 | 0.05405 | 11.6057 | 4.0e-07 | 0.50504 | 0.74959 |
| | | | | | Lower 99% | Upper 99% |
| Intercept | | | | | 0 | 0 |
| x1 | | | | | 0.45166 | 0.80298 |

107

# APPENDIX B: RAW DATA EXAMPLES

## A. DEPENDENT VARIABLE FILE

Team name: a
Experiment condition: 11111
True class: 1
Final assessement: 1
Final Score: 1
Final Confidence: 2
Time remaininig at final
decision: 13.000000
Number of leader's log entries: 6
Number of leader's queries: 0
Number of leader's opinion
changes: 1
Initial assessment of DM0: 2
Initial confidence of DM0: 1
Total number of probes: 35
Probe rate: 0.187166
Number of probes by DM1: 11
Number of probes by DM2: 10
Number of probes by DM3: 14
Number of log entries by DM1: 10
Number of log entries by DM2: 14
Number of log entries by DM3: 12
Total number of subordinates'
log: 36
Initial assessment of DM1: 2
Initial assessment of DM2: 1
Initial assessment of DM3: 1
Initial confidence of DM1: 1
Initial confidence of DM2: 1
Initial confidence of DM3: 1
Final assessment of DM1: 1
Final assessment of DM2: 1
Final assessment of DM3: 2
Final confidence of DM1: 2
Final confidence of DM2: 1
Final confidence of DM3: 2

Team name: a
Experiment condition: 11124
True class: 2
Final assessement: 2
Final Score: 1
Final Confidence: 2
Time remaininig at final
decision: 13.000000
Number of leader's log entries: 7
Number of leader's queries: 0
Number of leader's opinion
changes: 0
Initial assessment of DM0: 2
Initial confidence of DM0: 1
Total number of probes: 36
Probe rate: 0.192513
Number of probes by DM1: 12
Number of probes by DM2: 11
Number of probes by DM3: 13
Number of log entries by DM1: 10
Number of log entries by DM2: 12
Number of log entries by DM3: 12
Total number of subordinates'
log: 34
Initial assessment of DM1: 2
Initial assessment of DM2: 1
Initial assessment of DM3: 2
Initial confidence of DM1: 1
Initial confidence of DM2: 1
Initial confidence of DM3: 1
Final assessment of DM1: 2
Final assessment of DM2: 1
Final assessment of DM3: 2
Final confidence of DM1: 1
Final confidence of DM2: 1
Final confidence of DM3: 2

108

## B.    EVENT LOG FILE

filename: log11124.b

```
*
1 2010   12.000000
0 999
2.000000 0.000000 0.000000
201 1   10.000000 0.000000
1
*
2 2010   13.000000
1 999
2.000000 0.000000 0.000000
201 2   10.000000 0.000000
1
*
3 2010   15.000000
2 999
2.000000 0.000000 0.000000
201 3   10.000000 0.000000
1
*
2 2013   19.000000
   2 0 201 0 1
0.000000 0.000000 0.000000
2
*
1 2010   23.000000
0 999
2.000000 0.000000 0.000000
201 1   10.000000 0.000000
1
*
2 2010   25.000000
1 999
2.000000 0.000000 0.000000
201 2   10.000000 0.000000
1
*
3 2010   27.000000
2 999
2.000000 0.000000 0.000000
201 3   10.000000 0.000000
1
*
```

```
2 2013   32.000000
   2 0 201 1 3
0.000000 0.000000 0.000000
2
*
1 2010   35.000000
0 999
2.000000 0.000000 0.000000
201 1   10.000000 0.000000
1
*
3 2013   35.000000
   3 0 201 1 1
0.000000 0.000000 0.000000
2
*
2 2010   36.000000
1 999
2.000000 0.000000 0.000000
201 2   10.000000 0.000000
1
*
3 2010   39.000000
2 999
2.000000 0.000000 0.000000
201 3   10.000000 0.000000
1
*
0 2013   40.000000
   0 4 201 1 1
0.000000 0.000000 0.000000
1
```

# LIST OF REFERENCES

Egan, J.P., *Signal Detection Theory and ROC Analysis*, Academic Press, 1975.

Glass, A.L., Holyoak, K.J., and Santa, J.L, *Cognition*, Addison-Wesley, 1979.

Gough, M.J., *The Effects of Team Leader Feedback on Situation Assessment in Distributed Anti-Air Warfare Teams*, Master's Thesis, Navy Postgraduate School, Monterey, CA, March 1992.

Green, D.M, and Swets, J.A., *Signal Detection Theory and Psychophysics*, Wiley, 1974.

Kleinman, D.L, Serfaty, D., and Luh, P.B., "A Research Paradigm for Multi-Human Decision Making," in *Proc. American Contr. Conf.*, San Diego, CA 1984.

Kubovy, M., and Healy, A.F., "Process Models of Probabilistic Categorization." In Wallsten, T.S., (Ed.), *Cognitive Processes in Choice and Decision Behavior*, Lawrence Erlbaum Associates, Inc., 1980, pp. 239-262.

Lockhead, G.R., "Know, Then Decide." In Wallsten, T.S., (Ed.), *Cognitive Processes in Choice and Decision Behavior*, Lawrence Erlbaum Associates, Inc., 1980, pp. 143-154.

Macmillan, N.A., and Kaplan, H.L., "Detection Theory Analysis of Group Data: Estimating Sensitivity From Average Hit and False-Alarm Rates," *Psychological Bulletin*, Vol. 98, No. 1, pp. 185-199, 1985.

Mallubhatla, R., Pattipati, K.R., Kleinman, D.L., and Tang, Z., "A Model of Distributed Team Information Processing under Ambiguity," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 21, No. 4, pp. 713-725, July/August 1991.

Norman, D.A., "A Comparison of Data Obtained with Different False-Alarm Rates," *Psychological Review*, Vol. 71, No. 3, pp.243-246, 1964.

Payne, J.W., "Information Processing Theory: Some Concepts and Methods Applied to Decision Research." In Wallsten, T.S., (Ed.), *Cognitive Processes in Choice and Decision Behavior*, Lawrence Erlbaum Associates, Inc., 1980, pp. 95-116.

Pete, A., Pattipati, K.R., and Kleinman, D.L., "Distributed Detection in Teams with Partial Information: A Normative-Descriptive Model," *IEEE Transactions on Systems, Man, and Cybernetics*, 1993:1, in press.

Pete, A., Pattipati, K. R., and Kleinman, D. L., "Optimal Team and Individual Decision Rules in Uncertain Dichotomous Situations," *Public Choice*, 1993:2, in press.

Peterson, C.R. and Beach, L.R., "Man As an Intuitive Statistician," *Psychological Bulletin*, Vol. 63, No. 1, pp. 29-46, 1967.

Pitz, G.F., "The Very Guide of Life: The Use of Probabilistic Information for Making Decisions." In Wallsten, T.S., (Ed.), *Cognitive Processes in Choice and Decision Behavior*, Lawrence Erlbaum Associates, Inc., 1980, pp. 77-94.

Pratt, J.W. and Gibbons, J.D. *Concepts of Nonparametric Theory*, Springer-Verlag, 1981.

Snedecor, G.W. and Cochran, W.G., *Statistical Methods; 7th edition*, Iowa State University Press, 1980.

Swets, J.A., "Indices of Discrimination or Diagnostic Accuracy: Their ROCs and Implied Models," *Psychological Bulletin*, Vol. 99, No. 1, pp. 100-117, 1986:1.

Swets, J.A., "Form of Empirical ROCs in Discrimination and Diagnostic Tasks: Implications for Theory and Measurement of Performance," *Psychological Bulletin*, Vol. 99, No. 2, pp. 181-198, 1986:2.

Swets, J.A., "Measuring the Accuracy of Diagnostic Systems," *Science*, Vol. 240, No. 3, pp. 1285-1293, 1988.

Swets, J.A., Shipley, E.F., McKey, M.J., and Green, D.M., "Multiple Observations of Signals in Noise," *J. Acoust. Soc. Am.*, Vol. 31, pp. 514-52, 1959.

Tang, Z., Pattipati, K.R., and Kleinman, D.L., "An Algorithm for Determining the Decision Thresholds in a Distributed Detection Problem," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 21, No. 1, pp. 231-237, January/February 1991.

Tversky, A., and Kahneman, D., "Judgment under Uncertainty: Heuristics and Biases," *Science*, Vol. 185, pp. 1124-1131, 1974

# INITIAL DISTRIBUTION LIST

No. Copies

1. Defense Technical Information Center    2
   Cameron Station
   Alexandria VA 22304-6145

2. Library, Code 052    2
   Naval Postgraduate School
   Monterey CA 93943-5002

3. Naval Training Systems Center, Code 262    2
   12350 Research Parkway
   Orlando, FL 32826-3224

4. Professor Michael G. Sovereign, Code OR/SM    2
   Naval Postgraduate School
   Monterey, CA 93943

5. Professor William G. Kemple, Code OR/KE    1
   Naval Postgraduate School
   Monterey, CA 93943

6. David Kleinman    2
   University of Connecticut
   Department of Electrical & Systems Engineering
   Box U-157, Room 312
   260 Glenbrook Road
   Storrs, CT 06268

7. Dr. Elliot E. Entin    2
   ALPHATECH, INC.
   Executive Place III
   50 Mall Road
   Burlington, MA 01803

8.    Gerald S. Malecki                                          1
      Office of Naval Research
      Cognitive Neural Sciences Division
      Code 1142
      800 North Quincy Street
      Arlington, VA 22217

9.    Robert R. Armbruster, LT USN                               1
      Naval Submarine School
      Code 81, SOAC Class 93070
      Box 700
      Groton, CT 06349-5700